Covariate Selection for Generalizing Experimental Results[*]

Naoki Egami[†]   Erin Hartman[‡]

July 19, 2018

Abstract

Social and biomedical scientists are often interested in generalizing the average treatment effect (ATE) estimated in a randomized experiment to a non-experimental target population. Many existing approaches require observing a set of variables accounting for selection into the experiment (a *sampling* set) in the target population as well as in the experimental sample. To relax this strict data requirement, we propose a data-driven method to estimate a *separating* set – a set of variables affecting both the sampling mechanism and treatment effect heterogeneity. Our approach has two advantages. First, our algorithm only requires that a sampling set be observed in the experimental data, not in the target population. As long as researchers can collect a rich set of covariates on the experimental sample, the proposed method can select a separating set sufficient for identifying the population ATE. Second, we can incorporate researcher-specific data constraints. When scholars know certain variables are unmeasurable in the target population, our method can select a separating set subject to such constraints. We validate our proposed method using naturalistic simulations based on real-world data.

[*] We would like to thank Alexander Coppock, Chad Hazlett, and Don Green for their helpful comments, as well as the participants of the UCLA IDSS workshop and the Yale Quantitative Methods Workshop.

[†] Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: negami@princeton.edu, URL: http://scholar.princeton.edu/negami

[‡] Assistant Professor of Statistics and Political Science, University of California, Los Angeles, Los Angeles, CA 90095. Email: ekhartman@ucla.edu, URL: www.erinhartman.com

1 Introduction

Over the last few decades, social and biomedical scientists have developed and applied an array of statistical tools to make valid causal inference. In particular, randomized experiments have become the mainstay for estimating causal effects. Although many scholars agree upon the high internal validity of experimental results, there is still a debate about how scientists should infer the impact of policies and interventions on broader populations (Imai *et al.*, 2008; Angrist and Pischke, 2010; Imbens, 2010; Deaton and Cartwright, 2017). This issue of generalizability, also known as transportability, is pervasive in practice because randomized controlled trials are often conducted on non-representative samples (Cook *et al.*, 2002; Druckman *et al.*, 2011; Pearl and Bareinboim, 2014; Allcott, 2015; Stuart *et al.*, 2015). To generalize experimental results, we consider statistical methods to adjust for the relevant difference between an experimental sample and a target population.

Recent studies have formalized what types of variables researchers need to adjust for when generalizing experimental findings. Existing approaches adjust for one of the following three types of variables. (1) *a sampling set*: a set of variables explaining the sampling mechanism, i.e., how units are sampled into a given experiment (Cole and Stuart, 2010; Stuart *et al.*, 2011; Xie, 2013; Tipton, 2013; Pearl and Bareinboim, 2014; Hartman *et al.*, 2015; Buchanan *et al.*, 2018), (2) *a heterogeneity set*: a set of variables explaining treatment effect heterogeneity (Kern *et al.*, 2016; Nguyen *et al.*, 2017), and (3) *a separating set*: a set of variables affecting both the sampling mechanism and treatment effect heterogeneity (Kern *et al.*, 2016). Note that this separating set contains a sampling set and a heterogeneity set as special cases.

Despite these advances, empirical researchers have not yet widely taken up these methods. Previous approaches require rich covariate information in both the experimental sample and the target population, which might be unavailable in many applied settings (Stuart and Rhodes, 2016). In fact, most existing methods require a sampling set to be measured in the target population as well as in the experimental sample. Due to this strict data requirement, the issue of generalizability has often been intractable in practice, while applied researchers recognize its importance.

To overcome these challenges, we develop a new algorithm that can estimate a separating set from the experimental data. This algorithm only requires that a sampling set be observed in the experimental sample, not in the target population. We then show that researchers can in fact identify the population average treatment effect by adjusting for separating sets selected by the proposed algorithm.

Our approach has three advantages. First, unlike existing approaches relying on rich covariate information in both the experimental sample and the non-experimental target population, we only require rich covariate data in the experimental sample. This has practical implications because researchers often have more control over what to measure on their experimental subjects, even when it is difficult to collect detailed information about their target population. The second related advantage is that our covariate selection algorithm in fact only exploits the experimental data, not the target population data. Researchers can estimate a separating set before population data collection, which can inform which types of variables they should measure in the target population. Finally, we can also incorporate user-constraints on what variables can feasibly be collected in the target population. If there are characteristics that cannot be measured in the target population, the algorithm will identify a separating set subject to these constraints, if one is feasible.

Our article builds on a growing literature on the population average treatment effect, which has two general directions. First, most previous studies have focused on articulating identification assumptions (Cole and Stuart, 2010; Stuart *et al.*, 2011; Tipton, 2013; Pearl and Bareinboim, 2014; Hartman *et al.*, 2015; Buchanan *et al.*, 2018, e.g.). In particular, Kern *et al.* (2016) explicitly show that researchers have to consider treatment effect heterogeneity and the sampling mechanism jointly. Their assumption of strong ignorability for treatment effects, i.e. that the difference in potential outcomes is conditionally independent of sample selection, serves as a basis of our approach. However, to estimate the population effect, existing approaches often assume that researchers have access to a large number of covariates in both the experimental sample and the non-experimental target population. Our main contribution is to provide a new data-driven covariate selection algorithm to find a feasible separating set in situations where researchers have some data constraints in the target population.

Research in the second direction argues that the necessary assumptions for existing methods are often too strong in practice. Recent papers have explored methods for sensitivity analyses (Nguyen *et al.*, 2017; Andrews and Oster, 2017) and bounds (Chan, 2017) to achieve partial identification under weaker assumptions. Our paper is complementary to these approaches. We instead focus on the point identification of the population average treatment effect and alleviates strong assumptions about data requirements by adding an additional step of estimating a separating set.

2 Definition of Separating Sets

In this section, we first set up the potential outcomes notations and then define a separating set along with a sampling set and a heterogeneity set. This typology helps us clarify different research settings where we can identify the population average treatment effect.

2.1 The Setup

Let T_i be a binary treatment assignment variable for unit i with $T_i = 1$ for treatment and 0 for control. We define $Y_i(t)$ to be the potential outcome variable of unit i if the unit were to receive the treatment t for $t \in \{0, 1\}$. In this paper, we make a stability assumption, which states that there is neither interference between units nor different versions of the treatment (Cox, 1958; Rubin, 1990). We define pre-treatment covariates $\mathbf{X}_i$ to be any variables not affected by the treatment variable. Researchers are rarely able to conduct their experiment on the target population itself, and therefore often conduct an randomized experiment with a different population. Define a sampling indicator S_i taking 1 if unit i is in the experiment and 0 if unit i is in the target population. We assume that every unit has non-zero probability of being in the experiment and that pre-treatment variables have the common support in the experiment and the target population. With this notation, we consider cases in which the experimental sample and the target population don't overlap. Similar results hold for cases in which the experimental sample is a subset of the target population. Although units might be randomly sampled into the experiment, in many applications, they often select into the experiment, making the experimental sample non-representative.

We are interested in estimating the average treatment effect in the target population. We call this causal estimand the population average treatment effect (PATE) and define it formally as follows.

Definition 1 (Population Average Treatment Effect)

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0].$$

The treatment assignment mechanism is controlled by researchers within the experiment ($S_i = 1$). In contrast, for units in the target population ($S_i = 0$), it is unknown. Formally, we assume that the treatment assignment is randomized within the experiment.

Assumption 1 (Randomization in Experiment)

$$\{Y_i(1), Y_i(0), \mathbf{X}_i\} \perp\!\!\!\perp T_i \mid S_i = 1.$$

For each unit in the experimental condition, only one of the potential outcome variables can be observed, and the realized outcome variable for unit i is denoted by $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. While both outcomes and treatments need to be measured in the experimental data, we do not require information about treatments or outcomes for the non-experimental target population.

2.2 Separating Sets

Kern *et al.* (2016) show that the PATE can be identified by a set of variables affecting both treatment effect heterogeneity and the sampling mechanism (sample ignorability for treatment effects). In this paper, we refer to this set as a *separating set.* Formally, a separating set is any set that makes the sampling indicator and treatment effect heterogeneity conditionally independent.

Assumption 2 (Ignorability with A Separating Set (Kern *et al.*, 2016))
There exists a separating set $\mathbf{W}_i$ such that

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathbf{W}_i. \tag{1}$$

This definition of a separating set contains two simple cases: (1) when no treatment effect heterogeneity exists and (2) when the experimental sample is randomly drawn from the target population. In both of these cases, $\mathbf{W}_i = \{\varnothing\}$. This separating set also encompasses two common approaches in the literature as special cases. First, researchers often employ statistical methods based on a *sampling set* – a set of all variables affecting the sampling mechanism. Second, researchers might adjust for a *heterogeneity set* – a set of all variables governing treatment effect heterogeneity. Below, we formalize these sets based on potential outcomes.

We define a *sampling set* as a set of variables that determines the sampling mechanism by which individuals come to be in the experimental sample. For example, when a researcher implements a stratified sampling with gender and age, the sampling set consists of those two variables. When researchers control the sampling mechanism, a sampling set is known by design. However, when samples are selected without such an explicit sampling design, a sampling set is unknown and in practice, researchers must posit a sampling mechanism. Formally, we can define a sampling set $\mathbf{X}^S$ as follows.

$$\{Y_i(1), Y_i(0), \mathbf{X}_i^{-S}\} \quad \perp\!\!\!\perp \quad S_i \mid \mathbf{X}_i^S \tag{2}$$

where $\mathbf{X}_i^{-S}$ is a set of pre-treatment variables that are not in $\mathbf{X}_i^S$. This conditional independence means that the sampling set is a set that sufficiently explains the sampling mechanism. Given the sampling set, the sampling indicator is independent of the joint distribution of potential outcomes and all other pre-treatment covariates.[1] We refer to variables in the sampling set as sampling variables.

The other popular approach is to adjust for a set of all variables explaining treatment effect heterogeneity, which we call a *heterogeneity set*. Formally, we can define a heterogeneity set $\mathbf{X}^H$ as follows.

$$Y_i(1) - Y_i(0) \quad \perp\!\!\!\perp \quad \{S_i, \mathbf{X}_i^{-H}\} \mid \mathbf{X}_i^H, \tag{3}$$

where $\mathbf{X}_i^{-H}$ is a set of pre-treatment variables that are not in $\mathbf{X}_i^H$. In this case, because a heterogeneity set fully accounts for treatment heterogeneity, $Y_i(1) - Y_i(0)$ is independent of all other variables.

We want to emphasize that a sampling set and a heterogeneity set are special cases of a separating set. Yet, there may exist many different separating sets, which we explore in Section 3.

2.3 Identification of the Population Average Treatment Effect

In Section 2.2, we outlined three main conceptual types of sets: a sampling set, a heterogeneity set, and a separating set. Although often implicit in many empirical studies, it is critical to be explicit about what variables are measured in the experimental sample and the target population. For this reason, we distinguish between different research scenarios, or "settings", researchers may find themselves in depending on the available covariate information.

We begin with the most demanding scenario – when a separating set is known and measured in both the experimental sample and the target population. If this is the case, the PATE is nonparametrically identified.

[1] Note that one advantage here is that the sampling set yields conditional independence of the joint distribution of potential outcomes, rather than the difference, which allows for generalization of a wider range of estimands, not just the population average treatment effect defined as the difference in the potential outcomes.

Setting 1 (A Separating Set is Observed in Experiment and in Target Population)

A separating set $\mathbf{W}_i$ is observed in both the experimental sample ($S_i = 1$) and the target population ($S_i = 0$).

Theorem 1 (Identification of the PATE with Separating Sets)

In Setting 1, the PATE is identified with a separating set $\mathbf{W}_i$ under Assumptions 1 and 2.

$$\tau \;=\; \int \Big\{ \mathbb{E}[Y_i \mid T_i = 1, S_i = 1, \mathbf{W}_i = \mathbf{w}] - \mathbb{E}[Y_i \mid T_i = 0, S_i = 1, \mathbf{W}_i = \mathbf{w}] \Big\} dF_{\mathbf{W}_i \mid S_i = 0}(\mathbf{w}),$$

where $F_{\mathbf{W}_i \mid S_i = 0}(\mathbf{w})$ is the cumulative distribution function of $\mathbf{W}$ conditional on $S_i = 0$.

Recall that a sampling set and a heterogeneity set are special cases of a separating set. Therefore, when a sampling set $\mathbf{X}_i^S$ is observed in both the experimental sample and the target population, the PATE is identified under Assumptions 1 and 2. Similarly, we can also identify the PATE under the same set of assumptions when a heterogeneity set $\mathbf{X}_i^H$ is observed in both the experimental sample and the target population.

3 Identification of Separating Sets

The PATE is identified when we measure a separating set in both the experimental sample and target population. The advantage of this approach over existing approaches based on sampling and heterogeneity sets is that there may exist potentially many different separating sets and hence, researchers can choose a set subject to their data constraints. For example, researchers might want to measure as few variables as possible in the target population. Or researchers might already know they cannot measure, for instance, political knowledge in the target population.

In this section, we propose a data-driven method to select separating sets. First, we show that a separating set is estimable from the experimental data. In settings where both a sampling set and a heterogeneity set are observed in the experimental data, we can estimate an exact separating set. A key advantage of this result is that we only require the experimental data, not the target population data, to discover separating sets, should they exist.

In many applied research contexts, however, the heterogeneity set is not readily available even in the experimental data because it is inherently unobservable. The fundamental problem of causal inference (Holland, 1986) states that only one of two potential outcomes are observable, which implies that the causal effect is unobserved at unit level and so is the heterogeneity set. We therefore develop an additional method to find a variant of a separating set, which we call a *weak* separating set, just using knowledge of a sampling set. We show that a weak separating set can also be discovered exploiting only the experimental data, and it is sufficient for identifying the PATE. This set renders the sampling set conditionally independent of the observed outcome in the experimental data. Importantly, this approach requires measuring the sampling set in the experimental data, but not in the target population. Although this data requirement might be still stringent in some contexts, it is much weaker than the one necessary for widely-used existing approaches, such as those relying on a sampling set or a heterogeneity set. After deriving theoretical results on the identification of a separating set in this section, we provide an algorithm to estimate separating sets in the next section.

3.1 Identification of Separating Sets

We begin with settings in which a sampling set and a heterogeneity set are observed in the experimental sample. In this setting, we can use the experimental data to identify exact separating sets. Although this data requirement is still restrictive, we emphasize that it does not require rich data on the target population.

Setting 2 (Sampling and Heterogeneity Sets are Observed in Experiment)
A sampling set $\mathbf{X}^S$ and a heterogeneity set $\mathbf{X}^H$ are observed in the experiment ($S_i = 1$).

In this setting, a separating set is estimable as a set that makes the sampling set and the heterogeneity set conditionally independent within the experimental data.

Theorem 2 (Identification of Separating Sets in Experiment)

In Setting 2, for a set of pre-treatment variables $\mathbf{W}$, under Assumptions 1 and 2,

$$\widetilde{\mathbf{X}_i}^H \perp\!\!\!\perp \widetilde{\mathbf{X}_i}^S \mid \mathbf{W}_i, T_i, S_i = 1 \implies Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathbf{W}_i, \tag{4}$$

where $\widetilde{\mathbf{X}_i}^H$ and $\widetilde{\mathbf{X}_i}^S$ are the set difference $\mathbf{X}_i^H \setminus \mathbf{W}_i$ and $\mathbf{X}_i^S \setminus \mathbf{W}_i$, respectively.

We provide the proof in Appendix A. Theorem 2 states that as long as we can find a set that satisfies the testable conditional independence on the left hand side, the discovered set is guaranteed to be a separating set. That is, with a large enough sample size, we can find an exact separating set from the experimental data alone. An intuition behind this theorem has straightforward two steps. First, because a heterogeneity set $\mathbf{X}_i^H$ fully explains treatment effect heterogeneity $Y_i(1) - Y_i(0)$, S_i and $Y_i(1) - Y_i(0)$ are conditionally dependent only when S_i and $\mathbf{X}_i^H$ are conditionally dependent. In addition, because a sampling set $\mathbf{X}_i^S$ fully explains the sampling indicator S_i, S_i and $\mathbf{X}_i^H$ are conditionally dependent only when $\mathbf{X}_i^S$ and $\mathbf{X}_i^H$ are conditionally dependent. Taken together, S_i and $Y_i(1) - Y_i(0)$ are conditionally dependent only when $\mathbf{X}_i^S$ and $\mathbf{X}_i^H$ are conditionally dependent. Note that when $\mathbf{X}_i^H$ and $\mathbf{X}_i^S$ share some variables, those variables should always be in $\mathbf{W}_i$. While $\mathbf{X}_i^S$ and $\mathbf{X}_i^H$ are always possible separating sets that will meet this condition, this theorem implies that alternative sets that are smaller or that meet user's data restrictions might be identified. Using the selected separating set, researchers can identify the PATE based on Theorem 1.

3.2 Identification of Weak Separating Sets

While Theorem 2 allows us to discover separating sets using the experimental data, a key challenge would be to measure both a sampling set and a heterogeneity set in the experimental data. In particular, it is often difficult to measure the heterogeneity set in practice because it is inherently unobservable. In this subsection, we consider an alternative setting in which we need to measure only a sampling set in the experimental data.

While it is difficult to identify an exact separating set in this setting, we show that a modified version of a separating set – a *weak* separating set – is estimable from the experimental data under a much weaker assumption. We formally define a weak separating set as follows.

Assumption 3 (Ignorability with A Weak Separating Set)

There exists a weak separating set $\mathbf{W}$ such that

$$Y_i(t) \perp\!\!\!\perp S_i \mid \mathbf{W}_i \quad \text{for } t = \{0,1\}. \tag{5}$$

We refer to this as a *weak* separating set since it renders the marginal, not the joint, distribution of potential outcomes conditionally independent of the sampling mechanism. We turn now to our final setting researchers may find themselves in – that the sampling set is observed only in the experimental data. Previous work using the sampling set assumes it is measured in both the experimental sample and the target population. Since researchers often have much more control over what data is collected in the experiment, this final setting greatly relaxes the data requirements of the previous literature. When we measure the sampling set in the experimental data, we can identify a weak separating set as a set that makes the sampling set and the observed outcomes conditionally independent within the experimental data.

Setting 3 (A Sampling Set is Observed in Experiment)

A sampling set $\mathbf{X}^S$ is observed in the experimental sample ($S_i = 1$).

Theorem 3 (Identification of Weak Separating Sets in Experiment)

In Setting 3, for a set of pre-treatment variables $\mathbf{W}$, under Assumption 1 and 3,

$$Y_i \perp\!\!\!\perp \mathbf{X}_i^S \mid \mathbf{W}_i, T_i, S_i = 1 \implies Y_i(t) \perp\!\!\!\perp S_i \mid \mathbf{W}_i. \tag{6}$$

We provide the proof in Appendix A. Theorem 3 states that as long as we can find a set that makes the observed outcome Y conditionally independent of the sampling set within the experimental data, the discovered set is guaranteed to be a weak separating set. With a large enough sample size, we can find a weak separating set from the experimental data alone. An intuition behind this theorem is similar to the one used for Theorem 2. Because the sampling set $\mathbf{X}_i^S$ fully explains the sampling indicator S_i, if the sampling indicator S_i and the potential

Setting	**Data Requirements**		**Identification**
	Experiment	Target Population	
Setting 1	Separating set	Separating set	Theorem 1
(Special Case 1.1)	Sampling set	Sampling set	
(Special Case 1.2)	Heterogeneity set	Heterogeneity set	
Setting 2	Sampling Set Heterogeneity Set	User Specified Constraints	Theorems 1 and 2
Setting 3	Sampling set	User Specified Constraints	Theorems 3 and 4

Table 1: Identifying the PATE from randomized trials under different research settings depending on data requirements for the experimental sample and the target population. Note: Many previous approaches assume that a sampling set or a heterogeneity set is measured in both the experimental sample and the target population (Setting 1). Our approach (Settings 2 and 3) relaxes data requirements for the target population by introducing an additional step of estimating separating sets. Our algorithm introduced in Section 4 can select separating sets subject to user specified data constraints on the target population.

outcome $Y_i(t)$ are conditionally dependent, the sampling set $\mathbf{X}^S$ and the observed outcome Y_i are also conditionally dependent.

Once we have discovered a weak separating set using the experimental data, we can identify the PATE with this discovered set.

Theorem 4 (Identification of the PATE with Weak Separating Sets)

When a weak separating set $\mathbf{W}$ is observed both in the experimental sample and the target population, the PATE is identified with the weak separating set $\mathbf{W}$ under Assumptions 1 and 3.

$$\tau \;=\; \int \Bigg\{ \mathbb{E}[Y_i \mid T_i = 1, S_i = 1, \mathbf{W}_i = \mathbf{w}] - \mathbb{E}[Y_i \mid T_i = 0, S_i = 1, \mathbf{W}_i = \mathbf{w}] \Bigg\} dF_{\mathbf{W}_i \mid S_i = 0}(\mathbf{w}).$$

In Table 1, we summarize three research settings we have discussed. Previous approaches (Setting 1) assume a separating set is observed both in the experimental sample and the target population. Most common special cases are methods based on a sampling set (Special Case 1.1) or a heterogeneity set (Special Case 1.2). Although the identification of the PATE in this setting is straightforward (Theorem 1), it requires rich covariate information from the experimental sample and more importantly from the target population. Our approach relaxes data requirements for the target population by introducing an additional step of estimating separating sets. In Setting 2 where we observe both a sampling set and a heterogeneity set in the experimental sample, we can identify separating sets from the experimental data alone (Theorem 2). Setting 3 only requires observing a sampling set in the experimental sample and we can identify weak separating sets (Theorem 3). For both settings, the next section will introduce an algorithm that can select separating sets subject to user specified data constraints in the target population.

4 Estimation of Separating Sets and the PATE

In this section, we first propose an estimation algorithm to find a weak separating set. As shown in Theorem 3, the goal is to find a set that makes a sampling set and observed outcomes conditionally independent within the experimental data. We show how to apply Markov random fields to encode conditional relationships among observed covariates and then select a separating set. A similar algorithm can be used for finding an exact separating set. In Section 4.2, we discuss how to use discovered separating sets to estimate the PATE.

4.1 Estimation of Separating Sets

We start with a brief summary of our algorithm and then describe each step in order.

1. Specify all variables in a sampling set $\mathbf{X}^S$.
2. Estimate a Markov random field over outcomes, treatments and all observed pre-treatment covariates in the experimental sample.
3. Enumerate all simple paths[2] from Y to $\mathbf{X}^S$ in the estimated Markov graph.
4. Find sets that block all the simple paths from Y to $\mathbf{X}^S$ in the estimated Markov graph.

[2] A simple path is a path in a Markov graph that does not have repeating nodes.

Markov Random Fields: Review Theorem 3 implies that we can find a weak separating set by estimating a set of variables $\mathbf{W}$ that satisfies the following conditional independence.

$$Y_i \perp\!\!\!\perp \mathbf{X}_i^S \mid \mathbf{W}_i, T_i, S = 1 \tag{7}$$

To estimate this set, we employ an Markov random field (MRF). MRFs are statistical models that encode the conditional independence structure over random variables via graph separation rules. For example, suppose there are three random variables A, B and C. Then, $A \perp\!\!\!\perp B \mid C$ if there is no path connecting A and B when node C is removed from the graph (i.e., node C *separates* nodes A and B), so-called the global Markov property (Lauritzen, 1996). Using this general theory of MRFs, the estimation of a separating set can be recast as the problem of finding a set of covariates separating outcome variable Y and a sampling set $\mathbf{X}^S$ in an estimated Markov graph. Therefore, we can find a separating set that satisfies equation (7) as far as we can estimate the MRF over $\{Y, \mathbf{X}^S, T, \mathbf{Z}\}$ within the experimental data where we define $\mathbf{Z}$ to be all pre-treatment variables measured in the experimental data. Note that MRFs (also known as undirected graphical models) are used here to estimate conditional independence relationships between $\{Y, \mathbf{X}^S, T, \mathbf{Z}\}$ as an intermediate step of estimating a separating set. MRFs are not used to estimate the underlying causal directed acyclic graphs (DAGs).

We rely on a mixed graphical model (Yang *et al.*, 2015), which allows for both continuous and categorical variables. More concretely, we assume that each node can be modeled as the following exponential family distribution using the remaining variables.

$$\Pr(G_r \mid G_{-r}) = \exp\left\{\alpha_r G_r + \sum_{h \neq r} \theta_{r,h} G_r G_h + \varphi(G_r) - \Phi(G_{-r})\right\}, \tag{8}$$

where G_{-r} is a set of all random variables in a Markov graph except for variable G_r, base measure $\varphi(G_r)$ is given by the chosen exponential family, and $\Phi(G_{-r})$ is the normalization constant. For example, for a Bernoulli distribution, the conditional distribution can be seen as a logistic regression model.

$$\Pr(G_r \mid G_{-r}) = \frac{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h)}{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h) + 1}. \tag{9}$$

In general, we model each node using a generalized linear model conditional on the remaining variables. Using this setup, we can estimate the structure of the MRF by estimating parameters $\{\theta_{r,h}\}_{h \neq r}$; $\theta_{r,h} \neq 0$ for variable G_h in the neighbors of variable G_r and $\theta_{r,h} = 0$ otherwise. We estimate each generalized linear model with ℓ_1 penalty to encourage sparsity (Meinshausen and Bühlmann, 2006). Finally, using the AND rule, an edge is estimated to exist between variables G_r and G_h when $\theta_{r,h} \neq 0$ *and* $\theta_{h,r} \neq 0$. Researchers can also use an alternative OR rule (an edge exists when $\theta_{r,h} \neq 0$ *or* $\theta_{h,r} \neq 0$) and obtain the same theoretical guarantee of graph recovery.

Estimating Separating Sets Given the estimated graphical model, we can enumerate many different separating sets. First, we focus on the estimation of a separating set of the smallest size because it requires the smallest number of variables to be measured in the target population.[3] It is important to note that this separating set might not be the smallest with respect to the underlying DAG because MRFs don't encode all conditional independence relationships between variables. It is the smallest size among all separating sets estimable from MRFs.

We estimate this separating set from pre-treatment covariates $\mathbf{Z}$ as an optimization problem. A separating set should block all simple paths between outcome Y and variables in the sampling set $\mathbf{X}^S$. Therefore, we first enumerate all simple paths between Y and $\mathbf{X}^S$ and then find a minimum set of variables that intersect all paths.

Define q to denote the number of variables in $\mathbf{Z}$. We then define $\mathbf{d}$ to be a q-dimensional decision vector with d_j taking 1 if we include the j th variable of $\mathbf{Z}$ into a separating set and taking 0 otherwise. We use $\mathbf{P}$ to store all simple paths from Y to each variable in $\mathbf{X}^S$ where each row is a q-dimensional vector and its j th element takes 1 if the path contains the j th variable.[4]

[3] Other principled methods for choosing a set include precision in the PATE estimate or the costs associated with collecting target population information about the separating set.

[4] Simple paths can be enumerated using an off-the-shelf software, such as `igraph` in **R**.

With this setup, the estimation of this separating set of the smallest size is equivalent to the following linear programming problem given the estimated graphical model.

$$\min_{\mathbf{d}} \sum_{j=1}^{q} d_j$$
$$\text{s.t., } \mathbf{Pd} \geq \mathbf{1}.$$

where $\mathbf{1}$ is a vector of ones. The constraints above ensure that all simple paths intersect with at least one variable in a selected separating set, and the objective function just counts the total number of variables to be included into a separating set. Therefore, by optimizing this problem, we can find a set of variables with the smallest size that is guaranteed to block all simple paths.

It is important to emphasize that the estimation of the Markov graph is subject to uncertainty as any other statistical methods. Therefore, when the experimental data has small sample size, the estimation of the Markov graph might be too noisy to find an appropriate separating set. We investigate this issue of uncertainty in Section 5.

Incorporating Users' Constraints One advantage of our approach is that we can allow the flexibility for researchers to explicitly specify variables that they cannot measure in the target population. Thus, researchers can estimate a separating set such that it can be measured in both the experimental sample and the non-experimental target population. This is important in practice because it is often the case that researchers can measure a large number of covariates in the experimental data but they can collect relatively few variables in the target population. We can easily adjust the previous optimization problem to account for this restriction. Define $\mathbf{u}$ to be a q-dimensional vector with u_j taking 1 if we want to exclude the j th variable of $\mathbf{Z}$ from a separating set and taking 0 otherwise. Then, the optimization problem above changes as follows.

$$\min_{\mathbf{d}} \sum_{j=1}^{q} d_j$$
$$\text{s.t., } \mathbf{Pd} \geq \mathbf{1} \quad \text{and} \quad \mathbf{u}^\top \mathbf{d} = 0$$

Once we have estimated separating sets, we need to assess whether the conditional independence (equation (7)) holds in the experimental data. This exercise is similar to conducting a balance check after matching (Ho *et al.*, 2007), making sure that our algorithm achieves the desired conditional independence between observed outcomes and a sampling set.

In practice, it is possible that there exists no separating set, subject to user constraints. For example, the only separating set could consist of two variables, gender and political knowledge, and researchers specify political knowledge as a variable unmeasurable in the target population. In this case, there is no feasible separating set and our algorithm (correctly) finds no separating set.

4.2 Estimation of the Population Average Treatment Effect

Consistent with the two common approaches to the identification of the PATE, through the use of a sampling set or a heterogeneity set, the estimation of the PATE from the experimental data is done typically in one of two ways. Researchers either construct weights for adjusting the experimental data based on the sampling set, or they first model the treatment response surface based on the heterogeneity set and project predictions on to the target population (e.g., Nguyen *et al.*, 2017). Similar approaches can be taken with separating sets. Although both approaches can be combined with our algorithm of selecting separating sets, we discuss weighting approach as one example in this section.

The first step in estimation involves determining a set of weights for adjusting the experimental sample. In particular, we construct a set of calibration weights. Calibration is a general framework for weighting that minimizes the total distance, defined by some distance function $D(\cdot, \cdot)$, between an initial set of weights f_i and the final weights g_i subject to a set of moment constraints. It requires that population moments are equal to moments of the weighted experimental sample $\sum_{i;S_i=1} g_i \mathbf{W}_i$ (Deville and Särndal, 1992) where $\mathbf{W}_i$ is a separating set. Initial weights f_i are often set to be unity.

Common weighting techniques correspond to the use of different distance functions in the minimization of $\sum_{i;S_i=1} D(f_i, g_i)$. We calibrate on population means of variables in the separating set using raking, corresponding to a distance function $D(f_i, g_i) \propto g_i \times \log(g_i/f_i)$ (Deville *et al.*, 1993), which is the Kullback-Leibler divergence. Although here we use the marginal distributions, interactions between variables can also be incorporated into moments. Note that the additional parametric assumptions we impose for this estimation strategy alleviate the data requirements for researchers since the methods only require knowledge about population moments and not the full joint distributions of variables in the separating set. Whether these assumptions are warranted, or if alternative estimation techniques should be used, should be carefully considered in practice.

After constructing a set of weights, we estimate the individual or strata level effects. In particular, we use a generalized additive model (GAM) (Hastie and Tibshirani, 1990) to fit a fully interacted regression (Lin, 2013) to the weighed experimental sample. The PATE can then be estimated by taking the difference in the weighted mean predicted potential outcomes under treatment and control. Standard errors can be calculated using bootstrap, recalculating the calibration and GAM steps within each bootstrap replicate.

5 Simulation Studies

We turn now to simulations to explore how well the proposed algorithm can recover the PATE. We first verify that our proposed algorithm can obtain an unbiased estimate of the PATE. More importantly, we find that estimators based selected separating sets often have similar standard errors to the ones based on the true sampling set. Although our approach introduces an additional estimation step of finding separating sets to relax data requirements for the target population, it does not suffer from substantial efficiency loss. Both results hold with and without user constraints on variables measurable in the target population.

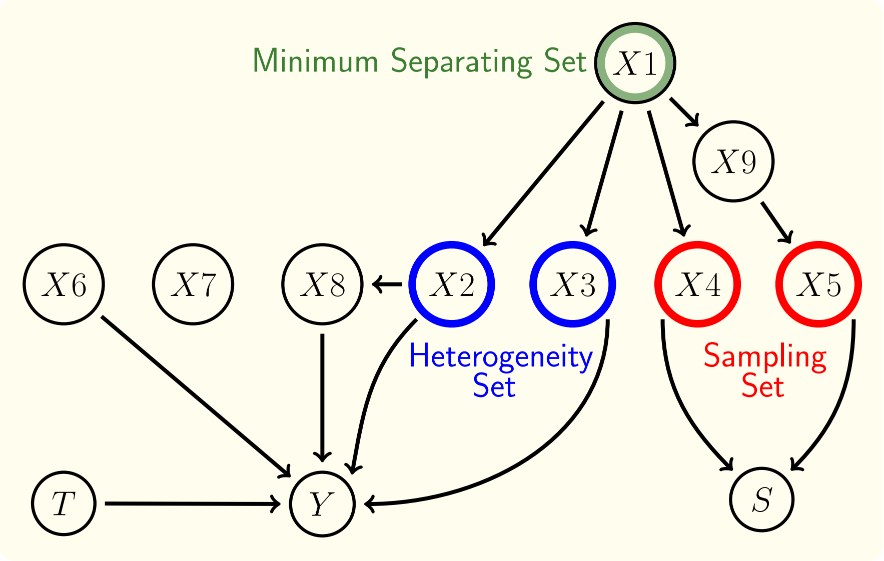

Figure 1: Causal DAG underlying the simulation study. Note: We consider three conceptually distinct sets (1) a sampling set, $X4$ and $X5$ (red), (2) a heterogeneity set, $X2$ and $X3$ (blue) and (3) the minimum separating set, $X1$ (green). Three root nodes $X1$, $X6$, $X7$ are normally distributed and other pre-treatment covariates are linear functions of their parents.

5.1 Simulation Design

In this subsection, we articulate our simulation design step by step. See Appendix B for all the details on the simulation design.

Pre-treatment Covariates and Potential Outcome Model To consider different types of separating sets, we assume the causal directed acyclic graph (DAG) in Figure 1 that encodes causal relationships among the outcome, the sampling indicator, and pre-treatment covariates. In this DAG, there are three conceptually distinct sets that we consider – (1) a sampling set, $X4$ and $X5$, depicted in red, (2) a heterogeneity set, $X2$ and $X3$, depicted in blue, and (3) the minimum separating set, $X1$, highlighted in green. Three root nodes $X1$, $X6$, $X7$ are normally distributed and other pre-treatment covariates are linear functions of their parents in the DAG.

The potential outcome is a linear function of its parents and an interaction between the treatment variable T and the heterogeneity set, $X2$ and $X3$. Two variables $X6$ and $X8$ only affect the potential outcomes under control, not causal effects. We draw $N \in$

$\{300, 600, 1500, 3000\}$ samples from this potential outcome model, which serve as the target population of this simulation. While the true PATE in each simulation varies depending on simulated target populations, it is set to 5.65 on average.

Sampling Mechanism and Treatment Assignment Out of this target population, we randomly sample a subset for a randomized experiment. The sampling mechanism is a probit model based on the sampling set, $X4$ and $X5$, with an average selection probability of 1/3. The treatment assignment mechanism is defined only for the experimental sample ($S_i = 1$). After being sampled into the experiment, every unit has the same probability of receiving the treatment $\Pr(T_i = 1 \mid S_i = 1) = 0.3$. For the sake of simplicity, we omit an arrow from the sampling indicator S to the treatment T in Figure 1.

Simulation Procedure Our simulation takes six steps in total.

Step 1: Generate N samples of pre-treatment covariates and potential outcomes where $N \in \{300, 600, 1500, 3000\}$.

Step 2: Randomly sample n units for experiments based on the sampling mechanism and assign treatments to them according to the treatment assignment mechanism. We use four different values as the size of experiments, $n = \{100, 200, 500, 1000\}$ corresponding to the N from Step 1.

Step 3: Estimate a weak separating set using the experimental data.

Step 4: Compute sampling weights so that the experimental data and the target population have the same mean values for each variable in the estimated separating set.

Step 5: Estimate the PATE. In particular, we run a fully interacted weighted GAM with the treatment variable and the estimated separating set. An estimate of the PATE is the difference between the weighted mean predicted potential outcomes under treatment and control.

Step 6: Compute bias and standard errors.

We repeat Step 1 to Step 6 for 4000 times. In each simulation, we also classify a type of an estimated separating set based on if it is a sampling set, a heterogeneity set, or the minimum separating set. By replacing Step 3 to Step 5, we can compare the performance of our proposed algorithm to four different estimators in terms of bias and standard errors: (1) a naive difference-in-means estimator, which does not adjust for any difference between the experimental sample and the target population, (2) an estimator based on an oracle sampling set, which skips Step 3 and adjusts for the true sampling set, (3) an estimator based on an oracle heterogeneity set, which skips Step 3 and adjusts for the true heterogeneity set, and (4) an estimator based on the oracle minimum separating set, which skips Step 3 and adjusts for the true minimum separating set.

5.2 Results

We present results in Figure 2. Not shown in the graph are the results for the naive difference-in-means, which has significant bias (0.35). As expected, we see that the bias is tends to zero for the oracle and estimated separating sets, and that the estimators are consistent for the PATE. More importantly, we see that estimators based the selected separating sets (red) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple). An estimator based on the heterogeneity set (orange) has significantly smaller standard errors than other estimators partly because it contains more of direct predictors of outcomes. However, this estimator might be unavailable in practice as discussed in Section 3 because a heterogeneity set is inherently unobservable.

Figure 3 shows the breakdown of types of estimated separating sets.[5] Since the algorithm can discover multiple separating sets, the frequency with which each set is chosen is presented, even though multiple sets can be chosen in any one simulation. We group sets that are

[5] Appendix B.2 presents numerical results for bias and standard error by selected type of separating set.

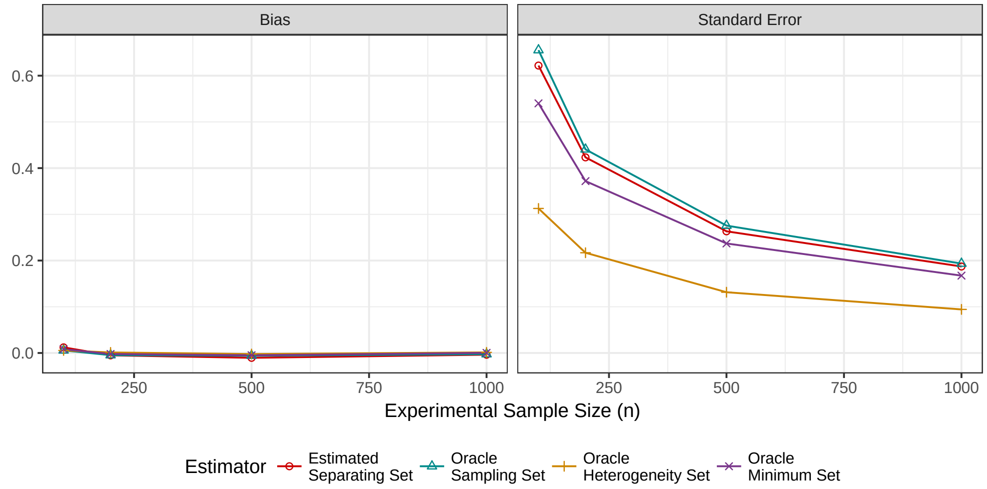

Figure 2: Simulation Results. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. More importantly, estimators based the estimated separating sets (red) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple). An estimator based on the heterogeneity set (orange) has significantly smaller standard errors than other estimators, but this estimator might be unavailable in practice.

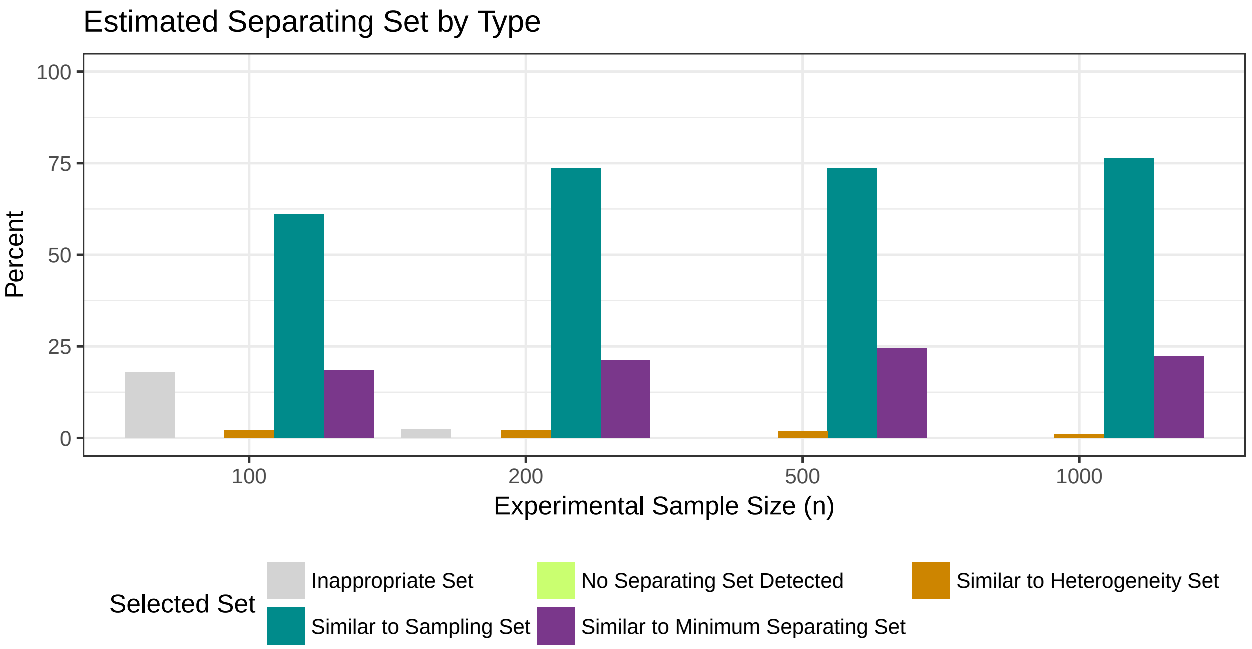

Figure 3: Types of Estimated Separating Sets. Note: We present the frequency of estimated separating sets by conceptual type. Over 75% of the time, the sampling set (green) is selected, and roughly one quarter of the time sets similar to the minimum separating set (purple) are selected. When the sample size is small, the proposed algorithm can select inappropriate sets (gray). However, the rate at which inappropriate sets are selected drops off rapidly with sample size.

conceptually similar. For example, if our algorithm selects the variables in the sampling set ($X4$ and $X5$) as well as an additional variable, we group these as "similar to" the sampling set. As can be seen, in these simulations, over 75% of the time, the sampling set (green) is selected, and roughly one quarter of the time something similar to the minimum separating set (purple) is selected. Small sample size can lead misestimation of the MRF, and therefore selection of inappropriate sets (gray) which do not remove bias – however the rate at which inappropriate sets are selected drops off rapidly with sample size.

5.3 Simulations Incorporating User Constraints

An advantage of our method is that researchers can specify variables that are unobservable in the target population. We conduct a simulation in which we specify that variable $X5$ is unobservable, thus making the sampling set unobservable in the target population. Figure 4 compares the estimated separating set which incorporates user constraints, to the oracle sampling and heterogeneity sets and the oracle minimum separating set. As theorems tell us, all sets let us estimate the PATE without bias.[6] Even with user constraints, we can verify that estimators based the estimated separating sets (red) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple).

Finally, we show the breakdown of types of estimated separating sets in Figure 5. Two patterns are worth noting. First, we see the algorithm selects each of the other types of separating sets more frequently. Even though the sampling set is not observable in this simulation, the algorithm often finds a set similar to the sampling set by using $X5$'s surrogate, $X9$. Second, as we expect, we observe more cases in which the proposed algorithm fails to find appropriate separating sets (around 10 %, light green).

[6] Results only include estimates for when the algorithm returned a separating set.

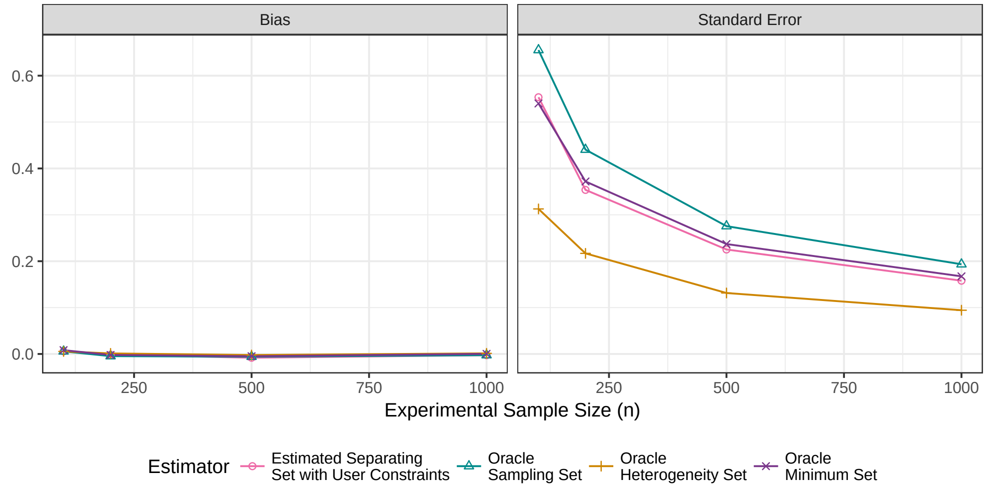

Figure 4: Simulation Results with User Constraints. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. In this simulation, we specify that variable $X5$ is unobservable, thus making the sampling set unobservable in the target population. Bias is close to zero for all estimators. Even with user constraints, estimators based the estimated separating sets (red) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple).

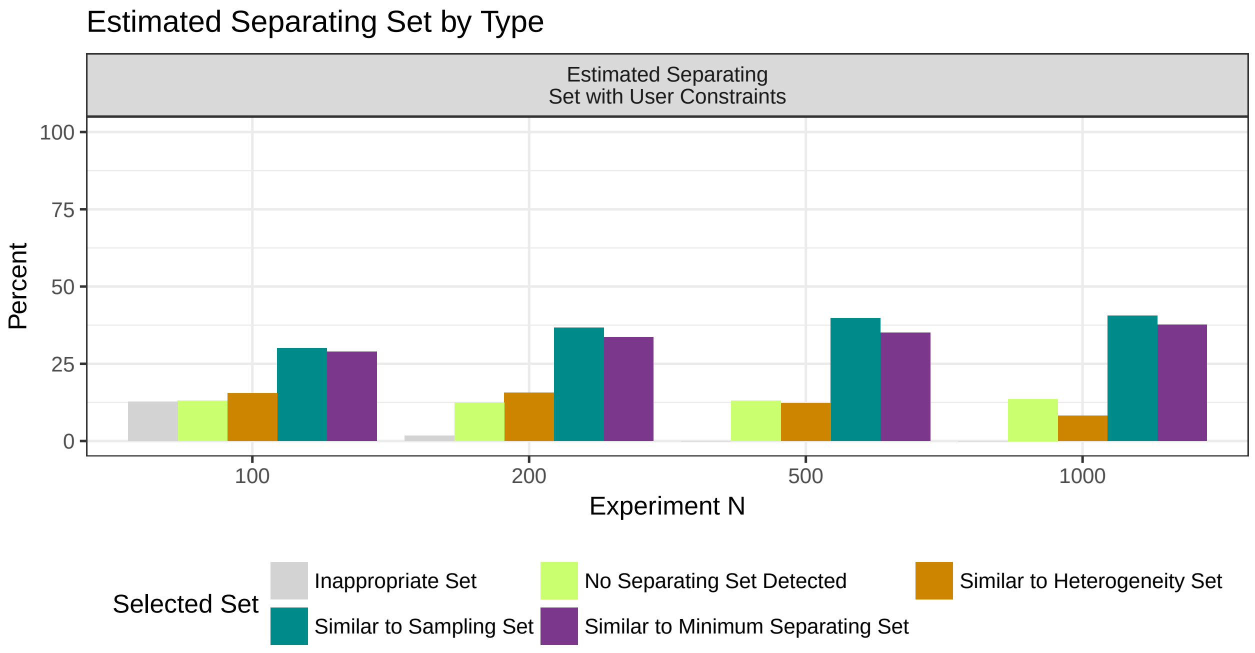

Figure 5: Type of Estimated Weak Separating Set with User Constraints. Note: We present the frequency of estimated separating sets by conceptual type. First, the algorithm selects each of the other types of separating sets more frequently. Second, as we expect, we observe more cases in which the proposed algorithm fails to find appropriate separating sets (around 10 %, light green

6 Validation with Naturalistic Simulation

In this section, we investigate the performance of our proposed approach using the real-world data from developmental economics. Banerjee *et al.* (2015) conducted randomized experiments to understand the effectiveness of a multifaceted development program in six countries – Ethiopia, Ghana, Honduras, India, Pakistan, and Peru. The program aims to foster self-employment among the ultra poor, and the researchers study the impact in the short and long term on a range of economic outcomes. In particular, they measure the causal impact of the program on 10 outcomes, ”consumption, food security, productive and household assets, financial inclusion, time use, income and revenues, physical health, mental health, political involvement, and womens empowerment” (Banerjee *et al.*, 2015, pg. 772).

While these experiments have high internal validity, generalizing the causal estimates to other populations is not straightforward for two reasons. First, although they find that the program is effective in all countries on most outcomes, they also show large treatment effect heterogeneity across different countries. This large heterogeneity is not surprising given that these countries have different baseline poverty levels and the programs were adjusted for each country's culture and economic context. Second, experimental units are not randomly sampled. These units are carefully targeted based on the needs: only the poorest households in the most economically disadvantaged regions of each country were eligible. Therefore, researchers need appropriate statistical adjustment techniques to generalize the causal findings across studies. Taken all together, this study offers us a unique situation to validate our proposed approach.

We investigate whether our method can recover the population average treatment by finding and then adjusting for a separating set. In particular, we are interested in two aspects: (1) Can our procedure recover the PATE by adjusting for a separating set? (2) Can we find a separating set smaller than a sampling set? To have clear answers to these questions, we design a naturalistic simulation based on the aforementioned real experiment. Unlike our simulations in Section 5 where we fully control the outcome model, we don't impose a simple parametric model for treatment heterogeneity. Instead, we mimic a realistic outcome model by estimating a nonparametric Bayesian model with the real data from Banerjee *et al.* (2015).

6.1 Simulation Design

In this subsection, we articulate our simulation design step by step.

Treatment and Outcome The treatment variable is binary, taking the value 1 if a household receives a development program and taking the value 0 otherwise.[7] Although they measure a range of variables at the household level, we use the index of financial inclusion as the main outcome of interest. The financial inclusion index includes measures related to borrowing power and savings. We choose this variable because the original study shows that it exhibits the largest treatment heterogeneity across countries, which makes generalization harder. To make the outcome variable comparable across countries, the authors standardize outcomes within each country using the control group. Therefore, the treatment effect of 0.1 corresponds to 0.1 standard deviation of the outcome variables in the control group.[8] We have a total of 9715 households. The sample size varies from 817 households (India) to 2525 households (Ghana).[9]

Potential Outcome Model and Heterogeneity Set To naturally simulate the potential outcome tables, we first estimate a Bayesian additive regression tree model (BART) (Chipman *et al.*, 2010; Hill, 2012; Green and Kern, 2012), regressing the main outcome variable on the treatment variable with country fixed effects and three pre-treatment variables related to financial inclusion. Hence, a heterogeneity set consists of country fixed effects and these three pre-treatment covariates. Using the fitted BART model, we predict the potential outcomes

[7] In some countries, they randomize treatments at the village level. See the original paper for details.

[8] See page 1260799-8 of the original paper.

[9] Our sample size differs from the number reported in the original paper due to missing data.

under both treatment and control for all units in the target population. We use these predicted potential outcomes as ”true” potential outcomes for this simulation.

Sampling Mechanism and Sampling Set In order to reflect realistic research settings, we non-randomly sample units for experiments. The sampling set includes a dummy variable for being in either Ethiopia or Ghana and three other variables: measures of the productive and household assets, consumption, and food security. Note that these three variables do not overlap with three variables used for the heterogeneity set. We design the sampling mechanism so that the mean probability of being in the experiment is about 65 %. We provide the exact sampling mechanism in Appendix C. To make generalization harder, we define the target population as units not sampled into the experiment. That is, the experimental sample and the target population are disjoint in this simulation. The average treatment effect in the experimental sample is about 0.16 and the population average treatment effect is about 0.58. Given that the estimate of the average treatment effect in the original study is 0.367 and the standard error is 0.03, this gap between the ATEs in the experimental sample and the target population is large.

Simulation Procedure Our simulation takes five steps. In the first step, we randomly add two variables from 20 pre-treatment variables available in the original data set. Hence, we assume that the algorithm has an access to a sampling set, a heterogeneity set, and these two variables as pre-treatment variables. In the second and third steps, we sample and then assign treatments to experimental units. In the fourth and fifth steps, we estimate and then adjust for a separating set to estimate the PATE. We describe each step in order.

Step 1: Randomly draw two auxiliary variables measured in the baseline survey.

Step 2: Sample experimental units according to the sampling mechanism.

Step 3: Randomly assign treatments within the experimental sample. We rely on complete randomization so that we can keep the total number of treated units fixed.

Step 4: Using the experimental data, we estimate an Markov graph over the outcome, the treatment, the heterogeneity set, the sampling set, and two auxiliary variables. Based on the estimated graph, we find separating sets. See Section 4.1.

Step 5: We weight experimental units so that the weighted experimental sample and the target population have the same mean values for each variable in the estimated separating set. We then estimate the PATE using the weighted experimental data. See Section 4.2.

We repeat Step 2 to Step 5 for 20 times in each of 100 simulations of Step 1. Therefore, in total, we estimate the PATE for 2000 simulations.

6.2 Results

The size of estimated separating sets are 4, 3, 2, and 1 with 74 %, 15 %, 7 %, and 4 %, respectively. Given that the size of the true sampling set is 4, this result means that our estimation algorithm finds separating sets as large as the sampling set 74 % of the time. Interestingly, the algorithm finds a separating set smaller than a sampling set in 26 % of the simulations. This simulation shows that discovered separating sets can be smaller than the sampling set with decently high probability.

Figure 6 reports how well our procedure can recover the PATE using the estimated separating set. The blue triangle is the mean of the PATE estimates based on the true sampling set. The blue bar shows the 95 % confidence interval for the PATE estimate. As known in the literature, it can correctly recover the true PATE. The rest of four estimates are from our procedure based on estimated separating sets. Separately for each size of the separating set, we show the mean and the 95 % confidence interval of the PATE estimate. Results based on the size 4 separating set show a similar distribution to the one from the sampling set. As shown in Section 5, an estimator based on selected separating sets have as short confidence intervals as the one based on the true sampling set. Finally, when the size of the estimated

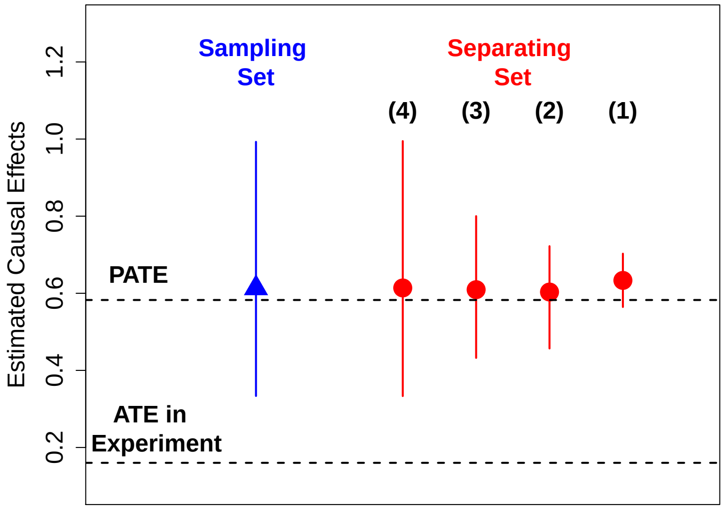

Figure 6: Estimates of the PATE from a sampling set and estimated separating sets. Note: Estimates based on the separating sets are separately reported based on the size of each set, which is indicated within parentheses. Results based on the size 4 separating set show a similar distribution to the one from the sampling set. When the size of the estimated separating set is smaller, the uncertainty around the PATE estimate is smaller. This is likely because weights based on the smaller number of variables are more stable.

separating set is smaller, the uncertainty around the PATE estimate is smaller. This is likely because weights based on the smaller number of variables are more stable, which results in smaller variance of the PATE estimates.

7 Concluding Remarks

The increased emphasis on well-identified causal effects in the social and biomedical sciences can sometimes lead researchers to narrow the focus of their research question and limit their findings to the experimental sample. However, primary research questions are often driven by the need to discover the impact of an intervention on a broader population. The extant literature has focused on the mathematical underpinnings concerning the generalizability of experimental evidence. The aim of this paper is to provide applied researchers with a means for uncovering a separating set using the experimental data alone.

Building on previous approaches to generalization, we clarify the role of the separating set – and its relationship to the sampling mechanism and treatment effect heterogeneity – in identification of population average treatment effects. This framework makes clear that there are many possible covariate sets researchers can use for the recovery of population effects, and it allows us to develop a new algorithm that can incorporate researchers' data constraints on the target population.

Through simulations, we verify that our proposed estimation technique – in which we find a separating set from the experimental data – can estimate the PATE without bias. More importantly, we provide evidence that estimators of the PATE based the selected separating sets can have similar standard errors to the ones based on the true sampling set under much weaker assumptions. We extend our findings to a naturalistic simulation, in which we use the data on six randomized trials conducted across three continents by Banerjee *et al.* (2015). In these naturalistic simulations we show that an estimated separating set can provide unbiased estimates of the PATE, and it can be smaller than the true sampling set in realistic applied settings.

Identifying population effects remains a challenging task for experimental researchers. The results here suggest researchers can increase a chance of generalization by collecting rich covariate information on their experimental subjects, even when their capacity of the population data collection is limited.

References

Allcott, H. (2015). Site selection bias in program evaluation. *Quarterly Journal of Economics*, pages 1117–1165.

Andrews, I. and Oster, E. (2017). Weighting for External Validity. *NBER working paper*.

Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, **24**(2), 3–30.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, **348**(6236), 1260799.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **95**, 1082.

Chan, W. (2017). Partially Identified Treatment Effects for Generalizability. *Journal of Research on Educational Effectiveness*, **10**(3), 646–669.

Chipman, H. A., George, E. I., McCulloch, R. E., *et al.* (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, **4**(1), 266–298.

Cole, S. R. and Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target PopulationsThe ACTG 320 Trial. *American journal of epidemiology*, **172**(1), 107–115.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2017). The generalizability of heterogeneous treatment effect estimates across samples. *Unpublished manuscript*.

Cox, D. R. (1958). *Planning of experiments.* Wiley.

Deaton, A. and Cartwright, N. (2017). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**(418), 376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, **88**(423), 1013–1020.

Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A. (2011). *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, **76**(3), 491–511.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **178**(3), 757–778.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models, volume 43 of monographs on statistics and applied probability.

Hill, J. L. (2012). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20**(1), 217–240.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, **15**(3), 199–236.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**(2), 481–502.

Imbens, G. W. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, **48**(2), 399–423.

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **179**(2), 319–376.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, **9**(1), 103–127.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedmans critique. *The Annals of Applied Statistics*, **7**(1), 295–318.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., and Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, **11**(1), 225–247.

Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, **29**(4), 579–595.

Rubin, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” [Ann. Agric. Sci. 10 (1923), 1–51]. *Statistical Science*, **5**(4), 472–480.

Stuart, E. A. and Rhodes, A. (2016). Generalizing Treatment Effect Estimates From Sample to Population: A Case Study in the Difficulties of Finding Sufficient Data:. *Evaluation Review*.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **174**(2), 369–386.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science*, **16**(3), 475–485.

Tipton, E. (2013). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, **38**(3), 239–266.

Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences*, **110**(16), 6262–6268.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, **16**(1), 3813–3847.

Appendix

A Proof of Theorems

Here, we provide proofs for all the theorems presented in the paper.

A.1 Proof of Theorem 1

$$\begin{aligned}
\tau &= \int_w \mathbb{E}[Y(1) - Y(0) \mid S = 0, \mathbf{W} = \mathbf{w}] dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \mathbb{E}[Y(1) - Y(0) \mid S = 1, \mathbf{W} = \mathbf{w}] dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y(1) \mid S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(0) \mid S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y(1) \mid T = 1, S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(0) \mid T = 0, S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y \mid T = 1, S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y \mid T = 0, S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}).
\end{aligned}$$

where the first equality follows from the rule of the conditional expectation, the second from the definition of a separating set (Assumption 2), the third from the linearity of the expectation, the fourth from Assumption 1, and the final follows from the consistency condition. □

A.2 Proof of Theorem 2

In this proof, we assume that the separating set $\mathbf{W}$ is disjoint with the sampling set $\mathbf{X}^S$ and the heterogeneity set $\mathbf{X}^H$ for simpler notations. The same proof applies to the case in which some variables of the sampling set or the heterogeneity set are in the separating set.

First, we have

$$\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, T, S = 1 \tag{10}$$

From Random Treatment Assignment(Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1 \tag{11}$$

Combining equations (10) and (11) (Contraction in Pearl (2009)),

$$\begin{aligned}
&\{\mathbf{X}^H, T\} \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1 \\
\Rightarrow\ &\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1.
\end{aligned} \tag{12}$$

Given that the conditional independence structure of $(\mathbf{X}^H, \mathbf{X}^S, \mathbf{Z})$ is the same under $S = 1$ and $S = 0$, (because S only changes the treatment assignment), we have

$$\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S \tag{13}$$

From the definition of the sampling variable,

$$\mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S \tag{14}$$

Combining equations (13) and (14) (Intersection in Pearl (2009)), we have

$$\begin{aligned} & \mathbf{X}^H \perp\!\!\!\perp \{S, \mathbf{X}^S\} \mid \mathbf{W} \\ \Rightarrow \quad & \mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}. \end{aligned} \tag{15}$$

Additionally, based on the definition of the heterogeneity set,

$$Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^H. \tag{16}$$

Therefore, by combining equations (15) and (16) based on Contraction in Pearl (2009),

$$\begin{aligned} & \{Y(1) - Y(0), \mathbf{X}^H\} \perp\!\!\!\perp S \mid \mathbf{W} \\ \Rightarrow \quad & Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}, \end{aligned}$$

which completes the proof. □

A.3 Proof of Theorem 3

First, we have

$$Y \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, T, S = 1 \tag{17}$$

From Random Treatment Assignment(Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1 \tag{18}$$

Combining equations (17) and (18) (Contraction in Pearl (2009)),

$$\begin{aligned} & \{Y, T\} \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1 \\ \Rightarrow \quad & Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \end{aligned} \tag{19}$$

Given that the conditional independence structure of $(Y(1), Y(0), \mathbf{X}^S, \mathbf{Z})$ is the same under $S = 1$ and $S = 0$, (because S only changes the treatment assignment, relationship for potential outcomes and pre-treatment variables would not change), we have

$$Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S \tag{20}$$

for $t = \{0, 1\}$.

From the definition of the sampling variable, for $t = \{0, 1\}$,

$$Y(t) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S \tag{21}$$

Combining equations (20) and (21) (Intersection in Pearl (2009)), we have

$$\begin{aligned} & Y(t) \perp\!\!\!\perp \{S, \mathbf{X}^S\} \mid \mathbf{W} \\ \Rightarrow \quad & Y(t) \perp\!\!\!\perp S \mid \mathbf{W} \end{aligned}$$

for $t = \{0, 1\}$. This completes the proof. □

A.4 Proof of Theorem 4

$$\begin{aligned}
\tau &= \int_w \mathbb{E}[Y(1) - Y(0) \mid S = 0, \mathbf{W} = \mathbf{w}] dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y(1) \mid S = 0, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(0) \mid S = 0, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y(1) \mid S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(0) \mid S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y(1) \mid T = 1, S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(0) \mid T = 0, S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}) \\
&= \int_w \Big\{ \mathbb{E}[Y \mid T = 1, S = 1, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y \mid T = 0, S = 1, \mathbf{W} = \mathbf{w}] \Big\} dF_{\mathbf{W}|S=0}(\mathbf{w}).
\end{aligned}$$

where the first equality follows from the rule of the conditional expectation, the second from the linearity of the expectation, the third from the definition of a weak separating set (Assumption 3), and the fourth from Assumption 1, and the final follows from the consistency condition. □

B Details on Simulation

B.1 Simulation Design

Pre-treatment Covariates We first generate the population using the following data generating process.

$$\begin{array}{ll}
X1 = 1/5 \times \epsilon_1 & X2 = 0.3 \times X1 + 0.5 \times \sqrt{1 - (0.6)^2} \times \epsilon_2 \\
X3 = 0.4 \times X1 + 0.5 \times \sqrt{1 - (0.8)^2} \times \epsilon_3 & X4 = 0.3 \times X1 + \sqrt{1 - (0.3)^2} \times \epsilon_4 \\
X5 = 0.8 \times X9 + \sqrt{1 - (0.8)^2} \times \epsilon_5 & X6 \sim N(2, 1) \\
X7 \sim N(3, 0.5) & X8 = (-0.3 \times X2 + \sqrt{1 - (0.3)^2} \times \epsilon_8)/4 + 2 \\
X9 = -0.4 \times X1 + \sqrt{1 - (0.4)^2} \times \epsilon_9 &
\end{array}$$

where auxiliary variables are defined as follow.

$$\begin{array}{llll}
\epsilon_1 \sim N(0, 1) & \epsilon_2 \sim N(2, 1) & \epsilon_3 \sim N(1, 1) & \epsilon_4 \sim N(-1, 1) \\
\epsilon_5 \sim N(0, 1) & \epsilon_8 \sim N(0, 1) & \epsilon_9 \sim N(0, 1) &
\end{array}$$

Potential Outcome Model We draw the potential outcomes as follows.

$$Y_i(t) = 4T + (6 \times X2 + 2 \times X3)T + 7 \times X2 + 8 \times X3 + X6 + 2 \times X8 + \epsilon_i$$

where $\epsilon_i \sim N(0, 0.1)$.

Sampling Mechanism We then draw a sampling indicator S_i as follows.

$$S_i = \Phi^{-1}\Bigg(\Phi(0.3) + (-0.5 \times X4 + 0.7 \times X5)/2.5 - \sum_{i=1}^{n} \{(-0.5 \times X4 + 0.7 \times X5)/2.5\}/n \Bigg),$$

where Φ is the normal cumulative distribution function and Φ^{-1} is the inverse of the normal cumulative distribution function. We draw samples proportional to this sampling weight and also ensure that the sample size of the experiment will be roughly 1/3 of the population.

B.2 Additional Simulation Results

Below is the bias and standard error result by selected estimated separating set type. We refer to sets that are "similar to" different conceptual sets in order to group sets that control for a specific type of separating sets, but which may include extra variables. For example, if the estimated set includes $X4$, $X5$, and $X8$, we say this is similar to a sampling set ($X4$ and $X5$). As theorems tell us, it doesn't matter what type of separating sets the algorithm estimates in the experimental data, all of them produce unbiased estimates so long as the set is an appropriate separating set (see Table 2). When an inappropriate set is chosen, which is common in the $n = 100$ case but rare as n increases, we see that inappropriate sets do not reduce bias. As we expect, when estimated separating sets are similar to a heterogeneity set, standard errors are the smallest.

		Similar to	Similar to	Similar to
Sample Size	Inappropriate Set	Heterogeneity Set	Separating Set	Sampling Set
100	0.1508	0.0624	0.0014	-0.0266
200	0.1229	0.0040	-0.0113	-0.0081
500	-0.1945	0.0088	-0.0057	-0.0124
1000	-0.0101	-0.0092	-0.0020	

Table 2: Simulation results for bias by type of estimated separating set.

		Similar to	Similar to	Similar to
Sample Size	Inappropriate Set	Heterogeneity Set	Separating Set	Sampling Set
100	0.5982	0.3082	0.5565	0.6503
200	0.3487	0.2131	0.3699	0.4440
500	0.1210	0.2308	0.2758	
1000	0.1039	0.1659	0.1938	

Table 3: Simulation results for standard error by type of estimated separating set.

C Details on Naturalistic Simulation

Sampling Mechanism

$$\text{logit}(\Pr(Y_i = 1 \mid \mathbf{X}_i^S)) = -2 + 2X_{i1}^S + 2X_{i2}^S + 2X_{i3}^S + 4X_{i4}^S$$

where

$X_{i1}^S =$ `Index of Review and Income`.

$X_{i2}^S =$ `Index of Consumption`.

$X_{i3}^S =$ `Index of Food Security`.

$X_{i4}^S = 1$ if a household is in Ethiopia or Ghana and $X_{i4}^S = 0$ otherwise.