# Online Supplementary Material

Covariate Selection for Generalizing Experimental Results

## SM-1 Proof of Theorems

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

### SM-1.1 Proof of Theorem 1

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

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

From Random Treatment Assignment(Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \tag{2}$$

Combining equations (1) and (2) (Contraction in Pearl (2000)),

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

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

$$\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S. \tag{3}$$

From the definition of the sampling variable,

$$\mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S. \tag{4}$$

Combining equations (3) and (4) (Intersection (Pearl, 2000)), we have

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

which implies

$$\mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}. \tag{5}$$

Additionally, based on the definition of the heterogeneity set,

$$Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^H. \tag{6}$$

Therefore, by combining equations (5) and (6) based on Contraction in Pearl (2000),

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

which implies $Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}$. □

### SM-1.2 Proof of Theorem 2

First, we have

$$Y \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, T, S = 1. \tag{7}$$

From Random Treatment Assignment(Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \tag{8}$$

Combining equations (7) and (8) (Contraction in Pearl (2000)),

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

which implies

$$Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \tag{9}$$

Given that the conditional independence structure of $(Y(1), Y(0), \mathbf{X}^S, \mathbf{W})$ 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

$$Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S, \tag{10}$$

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

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

$$Y(t) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S. \tag{11}$$

Combining equations (10) and (11) (Intersection in Pearl (2000)), we have

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

which implies

$$Y(t) \perp\!\!\!\perp S \mid \mathbf{W}$$

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

## SM-2 IPW Estimator

Here, we show that $\hat{\tau} \xrightarrow{p} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0]$.

**Proof** First, we rewrite the IPW estimator as follows.

$$\hat{\tau} = \frac{\frac{1}{n+m}\sum_i S_i \pi_i p_i T_i Y_i}{\frac{1}{n+m}\sum_i S_i \pi_i p_i T_i} - \frac{\frac{1}{n+m}\sum_i S_i \pi_i (1-p_i)(1-T_i) Y_i}{\frac{1}{n+m}\sum_i S_i \pi_i (1-p_i)(1-T_i)}, \tag{12}$$

where $n$ $(m)$ is the sample size of the experimental data (the population data). By the law of large number,

$$\begin{aligned}\frac{1}{n+m}\sum_i S_i \pi_i p_i T_i \xrightarrow{p} \mathbb{E}[S_i \pi_i p_i T_i] &= \mathbb{E}_{\mathbf{W}}\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) p_i \Pr(T_i = 1 \mid S_i = 1, \mathbf{W}_i)\} \\ &= \mathbb{E}_{\mathbf{W}}\left\{\frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)}\right\} = 1.\end{aligned}$$

Similarly, $\frac{1}{n+m}\sum_i S_i \pi_i (1-p_i)(1-T_i) \xrightarrow{p} 1$. Again, by the law of large number,

$$\frac{1}{n+m}\sum_i S_i \pi_i p_i T_i Y_i \xrightarrow{p} \mathbb{E}[S_i \pi_i p_i T_i Y_i], \quad \frac{1}{n+m}\sum_i S_i \pi_i (1-p_i)(1-T_i) Y_i \xrightarrow{p} \mathbb{E}[S_i \pi_i (1-p_i)(1-T_i) Y_i].$$

Hence, $\hat{\tau} \xrightarrow{p} \mathbb{E}[S_i \pi_i p_i T_i Y_i - S_i \pi_i (1-p_i)(1-T_i) Y_i]$. We focus on the term on the right.

$$\begin{aligned}
& \mathbb{E}\Bigg\{\pi_i\Bigg(S_i p_i T_i Y_i - S_i(1-p_i)(1-T_i)Y_i\Bigg)\Bigg\} = \mathbb{E}_{\mathbf{W}}\Bigg\{\pi_i \mathbb{E}\Bigg\{S_i p_i T_i Y_i - S_i(1-p_i)(1-T_i)Y_i \mid \mathbf{W}_i\Bigg\}\Bigg\} \\
= \; & \mathbb{E}\Bigg\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i)\mathbb{E}\Bigg\{p_i T_i Y_i - (1-p_i)(1-T_i)Y_i \mid S_i = 1, \mathbf{W}_i\Bigg\}\Bigg\} \\
= \; & \mathbb{E}\Bigg\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i)\{p_i \mathbb{E}[T_i Y_i \mid S_i = 1, \mathbf{W}_i] - (1-p_i)\mathbb{E}[(1-T_i)Y_i \mid S_i = 1, \mathbf{W}_i]\}\Bigg\} \\
= \; & \mathbb{E}\Bigg\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i)\Bigg(\mathbb{E}[Y_i(1) \mid S_i = 1, \mathbf{W}_i] - \mathbb{E}[Y_i(0) \mid S_i = 1, \mathbf{W}_i]\Bigg)\Bigg\} \\
= \; & \mathbb{E}\Bigg\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i)\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 1, \mathbf{W}_i]\Bigg\} = \mathbb{E}\Bigg\{\pi_i \Pr(S_i = 1 \mid \mathbf{W}_i)\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i]\Bigg\} \\
= \; & \mathbb{E}\Bigg\{\frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)}\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i]\Bigg\} \\
= \; & \int_{\mathbf{W}}\Bigg\{\frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)}\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i]\Bigg\}p(\mathbf{W})d\mathbf{W} \\
= \; & \int_{\mathbf{W}} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i]p(\mathbf{W} \mid S_i = 0)d\mathbf{W} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0],
\end{aligned}$$

where the first equality follows from the law of conditional expectation given $\mathbf{W}$, the second from the conditional expectation given $S$, the third from the linearity of expectation, the fourth from the conditional expectation given $T$, the fifth from the linearity of expectation, the sixth from the definition of separating $\mathbf{W}$, the seventh from the definition of $\pi$, the eight from the rule of expectation, the ninth from Bayes rule, and the tenth from the rule of expectation.

## SM-3 Markov Random Fields: Review

A Markov random field (MRF), also known as an undirected graphical model, is a popular statistical model that encodes the conditional independence structure over multiple observed random variables. The main advantage of the MRF is that it encodes the conditional independence relationships of many random variables compactly. While many important results have been derived for MRFs, we focus on one key property, so-called, the global Markov property, which we use in our paper.

MRFs define the conditional independence relationships via simple graph separation rules (Lauritzen, 1996). For sets of nodes $A$, $B$, and $C$, $A \perp\!\!\!\perp B \mid C$ if and only if there is no path connecting $A$ and $B$ when nodes in $C$ are removed from the graph (i.e., nodes in $C$ separates nodes $A$ and $B$). For example, in Figure SM-1, suppose $A = \{V_1, V_2, V_3\}$ and $B = \{V_6, V_7\}$. Then, if we define $C = \{V_4, V_5\}$, there is no path connecting $A$ and $B$ once nodes in $C$ are removed from the graph. Therefore, Figure SM-1 encodes the conditional independence relationship, $\{V_1, V_2, V_3\} \perp\!\!\!\perp \{V_6, V_7\} \mid V_4, V_5$.

As emphasized in the paper, we use the MRF as the statistical model to characterize the conditional independence relationships between observed random variables. We do not use the MRF as a step to estimate the underlying causal DAG.

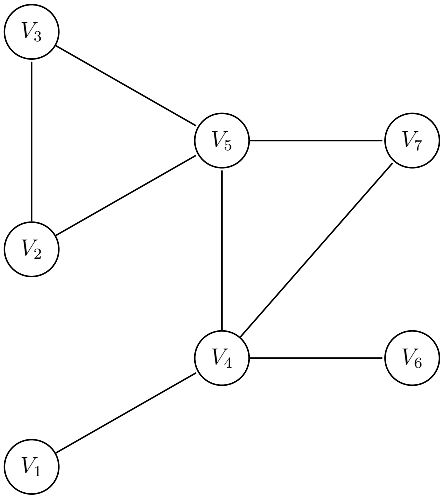

Figure SM-1: Example of a Markov Random Field (MRF).

## SM-4 Additional Results on Empirical Analysis

In Section 5, we focused on the inverse probability weighting estimator (equation (10)) to maintain the clear comparison with the original analysis that uses the weighting approach. In this section, we report results based on an outcome-model-based estimator[1] and a doubly robust estimator (Dahabreh *et al.*, 2019). In particular, for the outcome-model-based estimator, we use a fully-interacted linear model. Within the experimental data, we estimate a linear regression with a specified set of covariates separately for the treatment and control groups. Then, we use the estimated models to predict potential outcomes under treatment and control for the target population data. This outcome-model-based estimator is consistent under the assumption that the outcome model is correctly specified. For the doubly robust estimator, we use an augmented IPW estimator (Robins *et al.*, 1994; Dahabreh *et al.*, 2019) where the outcome model is a fully-interacted linear model and the sampling model is a logistic regression specified in Section 4.3. This doubly robust estimator is consistent if one of the two models — outcome or sampling models — is correctly specified.

We first extend our analyses in Section 5.1. Table SM-1 reports results based on the outcome-model-based estimator (an extension of Table 3). Similarly to the case of the IPW estimator, we find that (1) point estimates based on estimated separating sets are similar to those based on the original sampling set, and (2) standard errors based on our proposed estimated separating sets are smaller for 16 out of 17 outcomes. Table SM-2 reports results based on the doubly robust estimator (an extension of Table 3). Similarly to the cases of the IPW estimator and the outcome-model-based estimator, we find that (1) point estimates based on estimated separating sets are similar to those based on the original sampling set, and (2) standard errors based on our proposed estimated separating sets are smaller for 15 out of 17 outcomes. Therefore, for all three classes of estimators, our proposed approach of using the estimated separating set improves estimation accuracy. Finally, we also compare estimates across three classes of estimators in Table SM-3. Across 17 outcomes, we find that estimates of the PATE are relatively stable across different estimators (none of the differences in estimates are statistically significant at the conventional 0.05 level), which suggests model misspecification is of little concern.

We next extend our analyses in Section 5.2. Table SM-4 reports results for Section 5.2 by comparing estimates from the outcome-model-based estimator and the doubly robust estimator

[1] For outcome-model-based estimators, it is unclear whether adjusting for a smaller set of covariates leads to an increase in estimation efficiency; it will depend on how predictive are those covariates. However, at least in our application, we see below in Table SM-1 that outcome-model-based estimators based on estimated separating sets have smaller standard errors than those based on the original sampling set for 16 out of 17 outcomes. For outcome-model-based estimators, another benefit of having a smaller valid separating set is that it is easier for analysts to model the conditional expectation correctly with a fewer variables — the key necessary assumption for outcome-model-based estimators. We leave further technical and thorough investigation of outcome-model-based estimators for future work.

to estimates from the IPW estimator. While the point estimate for "Agricultural" is unstable due to a relatively large standard error (the first row in Table SM-4), estimates of the PATE are relatively stable across different estimators (none of the differences in estimates are statistically significant at the conventional 0.05 level), which again suggests model misspecification is of little concern.

| | Original Sampling Set | | Estimated Separating Set | |
|---|---|---|---|---|
| | Estimate | S.E. | Estimate | S.E. |
| Average employment hours | 4.58 | 2.35 | 3.57 | 1.80 |
| Agricultural | -0.00 | 1.61 | -1.22 | 1.45 |
| Nonagricultural | 4.58 | 1.77 | 4.79 | 1.45 |
| Skilled trades only | 3.70 | 1.03 | 4.08 | 0.86 |
| No employment hours | -0.04 | 0.03 | -0.03 | 0.02 |
| Any skilled trade | 0.27 | 0.05 | 0.25 | 0.04 |
| Works mostly in a skilled trade | 0.02 | 0.02 | 0.04 | 0.02 |
| Cash earnings | 5.20 | 7.31 | 8.22 | 7.02 |
| Durable assets | 0.08 | 0.10 | 0.06 | 0.08 |
| Vocational training | 0.52 | 0.05 | 0.50 | 0.04 |
| Hours of vocational training | 250.32 | 34.71 | 280.24 | 27.43 |
| Business assets | 340.79 | 141.74 | 367.61 | 127.23 |
| Maintain records | 0.14 | 0.05 | 0.14 | 0.04 |
| Registered | 0.03 | 0.04 | 0.04 | 0.03 |
| Pays taxes | 0.01 | 0.05 | 0.02 | 0.05 |
| Changed parish | 0.04 | 0.06 | -0.02 | 0.03 |
| Lives in Urban area | -0.01 | 0.03 | -0.01 | 0.03 |

Table SM-1: Estimates of the PATEs based on Outcome-Model-Based Estimator, comparing the Original Sampling Set and Estimated Exact Separating Sets. Extension of Table 3 in Section 5.1.

| | Original Sampling Set | | Estimated Separating Set | |
|---|---|---|---|---|
| | Estimate | S.E. | Estimate | S.E. |
| Average employment hours | 2.49 | 3.16 | 2.48 | 2.73 |
| Agricultural | -2.27 | 2.73 | -2.08 | 1.83 |
| Nonagricultural | 5.10 | 2.99 | 4.56 | 2.24 |
| Skilled trades only | 2.29 | 1.66 | 3.71 | 1.14 |
| No employment hours | 0.01 | 0.03 | -0.01 | 0.03 |
| Any skilled trade | 0.24 | 0.07 | 0.24 | 0.06 |
| Works mostly in a skilled trade | -0.03 | 0.03 | 0.02 | 0.04 |
| Cash earnings | 4.16 | 8.06 | 9.02 | 7.54 |
| Durable assets | 0.02 | 0.15 | 0.14 | 0.15 |
| Vocational training | 0.49 | 0.07 | 0.50 | 0.05 |
| Hours of vocational training | 228.35 | 50.14 | 283.26 | 34.55 |
| Business assets | 326.58 | 178.44 | 371.18 | 139.65 |
| Maintain records | 0.16 | 0.07 | 0.17 | 0.07 |
| Registered | 0.05 | 0.06 | 0.06 | 0.05 |
| Pays taxes | -0.02 | 0.07 | 0.01 | 0.07 |
| Changed parish | 0.06 | 0.07 | -0.04 | 0.05 |
| Lives in Urban area | 0.01 | 0.05 | -0.01 | 0.04 |

Table SM-2: Estimates of the PATEs based on Doubly Robust Estimator, comparing the Original Sampling Set and Estimated Exact Separating Sets. Extension of Table 3 in Section 5.1.

|  | IPW<br>Estimator |  | Outcome-Model-based<br>Estimator |  | AIPW<br>Estimator |  |
|---|---|---|---|---|---|---|
|  | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| Average employment hours | 4.79 | 2.39 | 3.57 | 1.80 | 2.48 | 2.73 |
| Agricultural | 0.30 | 1.69 | -1.22 | 1.45 | -2.08 | 1.83 |
| Nonagricultural | 4.49 | 1.79 | 4.79 | 1.45 | 4.56 | 2.24 |
| Skilled trades only | 4.36 | 0.99 | 4.08 | 0.86 | 3.71 | 1.14 |
| No employment hours | -0.03 | 0.03 | -0.03 | 0.02 | -0.01 | 0.03 |
| Any skilled trade | 0.27 | 0.06 | 0.25 | 0.04 | 0.24 | 0.06 |
| Works mostly in a skilled trade | 0.04 | 0.03 | 0.04 | 0.02 | 0.02 | 0.04 |
| Cash earnings | 12.54 | 5.11 | 8.22 | 7.02 | 9.02 | 7.54 |
| Durable assets | 0.18 | 0.13 | 0.06 | 0.08 | 0.14 | 0.15 |
| Vocational training | 0.53 | 0.05 | 0.50 | 0.04 | 0.50 | 0.05 |
| Hours of vocational training | 337.59 | 40.77 | 280.24 | 27.43 | 283.26 | 34.55 |
| Business assets | 425.02 | 135.65 | 367.61 | 127.23 | 371.18 | 139.65 |
| Maintain records | 0.20 | 0.07 | 0.14 | 0.04 | 0.17 | 0.07 |
| Registered | 0.09 | 0.05 | 0.04 | 0.03 | 0.06 | 0.05 |
| Pays taxes | 0.05 | 0.05 | 0.02 | 0.05 | 0.01 | 0.07 |
| Changed parish | -0.01 | 0.04 | -0.02 | 0.03 | -0.04 | 0.05 |
| Lives in Urban area | -0.01 | 0.04 | -0.01 | 0.03 | -0.01 | 0.04 |

Table SM-3: Estimates of the PATEs based on Estimated Exact Separating Sets for Three Estimators. Extension of Section 5.1.

|  | IPW<br>Estimator |  | Outcome-Model-based<br>Estimator |  | AIPW<br>Estimator |  |
|---|---|---|---|---|---|---|
|  | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| Agricultural | 0.64 | 1.63 | -1.10 | 1.31 | -1.32 | 1.56 |
| Changed parish | 0.05 | 0.03 | 0.03 | 0.03 | 0.04 | 0.03 |
| Lives in Urban area | -0.02 | 0.03 | -0.02 | 0.03 | -0.01 | 0.04 |

Table SM-4: Estimates of the PATEs based on Estimated Marginal Separating Sets for Three Estimators. Extension of Section 5.2.

## SM-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 a consistent estimator of the PATE. More importantly, we find that estimators based on estimated 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 what variables can be measured in the target population.

### SM-5.1 Simulation Design

In this subsection, we articulate our simulation design step by step. See the supplementary material 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 SM-2 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 green, (2) a heterogeneity set, $X2$ and $X3$, depicted in orange, and (3) the minimum separating set, $X1$, highlighted in purple. Three root nodes $X1$, $X6$, $X7$ are normally distributed and other pre-treatment covariates are linear functions of their parents in the DAG. In particular, pre-treatment covariates are generated as follows.

$$\begin{aligned}
&X1 \sim \mathcal{N}(0,1)\\
&X2 = 0.7 \times X1 + \sqrt{1-0.7^2} \times \epsilon_2\\
&X3 = 0.7 \times X1 + \sqrt{1-0.7^2} \times \epsilon_3\\
&X4 = 0.7 \times X1 + \sqrt{1-0.7^2} \times \epsilon_4\\
&X5 = 0.3 \times X9 + \sqrt{1-0.3^2} \times \epsilon_5\\
&X6 \sim \mathcal{N}(0,1)\\
&X7 \sim \mathcal{N}(0,1)\\
&X8 = -0.7 \times X2 + \sqrt{1-0.7^2} \times \epsilon_8\\
&X9 = 0.6 \times X1 + \sqrt{1-0.6^2} \times \epsilon_9
\end{aligned}$$

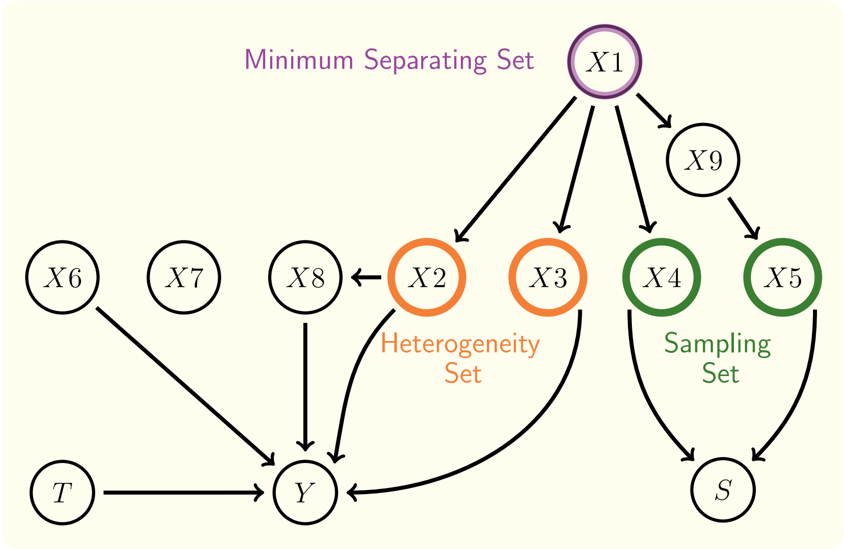

Figure SM-2: Causal DAG underlying the simulation study. Note: We consider three conceptually distinct sets (1) a sampling set, $X4$ and $X5$ (green), (2) a heterogeneity set, $X2$ and $X3$ (orange) and (3) the minimum separating set, $X1$ (purple). Three root nodes $X1$, $X6$, $X7$ are normally distributed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

where $\epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_8, \epsilon_9$ are drawn independently and identically from a standard normal distribution, $\mathcal{N}(0, 1)$. This results in the following correlation structure for variables $X1 - X9$.

$$cor(\mathbf{X}) = \begin{pmatrix} 1.00 & -0.70 & 0.70 & 0.70 & -0.20 & 0.00 & 0.00 & 0.50 & -0.70 \\ -0.70 & 1.00 & -0.50 & -0.50 & 0.15 & 0.00 & 0.00 & -0.70 & 0.50 \\ 0.70 & -0.50 & 1.00 & 0.50 & -0.15 & 0.00 & 0.00 & 0.33 & -0.50 \\ 0.70 & -0.50 & 0.50 & 1.00 & -0.15 & 0.00 & 0.00 & 0.33 & -0.50 \\ -0.21 & 0.15 & -0.15 & -0.15 & 1.00 & 0.00 & 0.00 & -0.10 & 0.30 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.50 & -0.70 & 0.33 & 0.33 & -0.10 & 0.00 & 0.00 & 1.00 & -0.33 \\ -0.70 & 0.50 & -0.50 & -0.50 & 0.30 & 0.00 & 0.00 & -0.33 & 1.00 \end{pmatrix}$$

We then draw the potential outcomes as follows.

$$Y_i(T_i) = 5T_i + 10 \times X_{3i} \times T_i - 10 \times X_{2i} \times T_i + X_{6i} - 3 \times X_{8i} + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$. Thus, the true PATE is set to 5.

**Sampling Mechanism and Treatment Assignment.** We randomly sample a set of $n$ units for a randomized experiment. The sampling mechanism is a logit model based on the sampling set, $X4$ and $X5$. 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 $\Pr(T_i = 1 \mid S_i = 1) = 0.5$. For the sake of simplicity, we omit an arrow from the sampling indicator $S$ to the treatment $T$ in Figure 1.

In particular, we draw a sampling indicator $S_i$ as follows. The second step scales the probability to be bounded away from zero and one.

$$\begin{aligned} S'_{i,lp} &= -20 \times X_{4i} + 20 \times X_{5i} \\ S_{i,lp} &= 0.25(S'_{i,lp} - \overline{S'_{lp}})/sd(S'_{lp}) \\ S_i &= \frac{1}{1 + e^{-S_{i,lp}}} \end{aligned}$$

**Simulation Procedure.** We conduct 5000 simulations for each of six experimental sample sizes, $n = \{100, 200, 500, 1000, 2000, 3000\}$. Within each simulation, we first randomly sample $n$ units for the experiment based on the sampling mechanism and randomly assign units to treatment according to the specified treatment assignment mechanism. We also randomly sample a target population of size $m = 10000$. We then estimate both an exact and a marginal separating set using the experimental data. An advantage of our method is that researchers can specify variables that cannot be measured in the target population. To illustrate this benefit, we also estimate a marginal separating set with a constraint that variable $X1$ is unmeasurable in the target population, thus making the minimal separating set unobservable in the target population. We compare these sets to an oracle sampling set, oracle heterogeneity set, and oracle minimum separating set.

For each estimated and oracle set, we compute the PATE using the inverse probability weighting estimator described in Section 4.3. In the supplementary material, we repeat these simulations with a calibration estimator discussed in Hartman *et al.* (2015), and a linear regression projection estimator.

## SM-5.2 Results

We present results in Figure SM-3. Not shown in the graph are the results for the naive difference-in-means, which has significant bias ($-1.0$). As expected, we see that the bias goes 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 on the selected marginal separating sets (red), exact separating sets (dark blue), and marginal separating set with user constraints (pink) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple). An estimator based on the oracle heterogeneity set (orange) has smaller standard errors than other estimators partly because it contains variables which are direct predictors of outcomes.

Figure SM-4 shows the breakdown of types of estimated separating sets. We group sets that are conceptually similar, and the frequency with which each set is chosen is presented. 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 as $n$ gets large, over 75% of the time, the minimal separating set (purple) is selected. Small sample size can lead to the misestimation of the MRF, and therefore selection of inappropriate sets (gray) which do not remove bias — however, the rate at which inappropriate

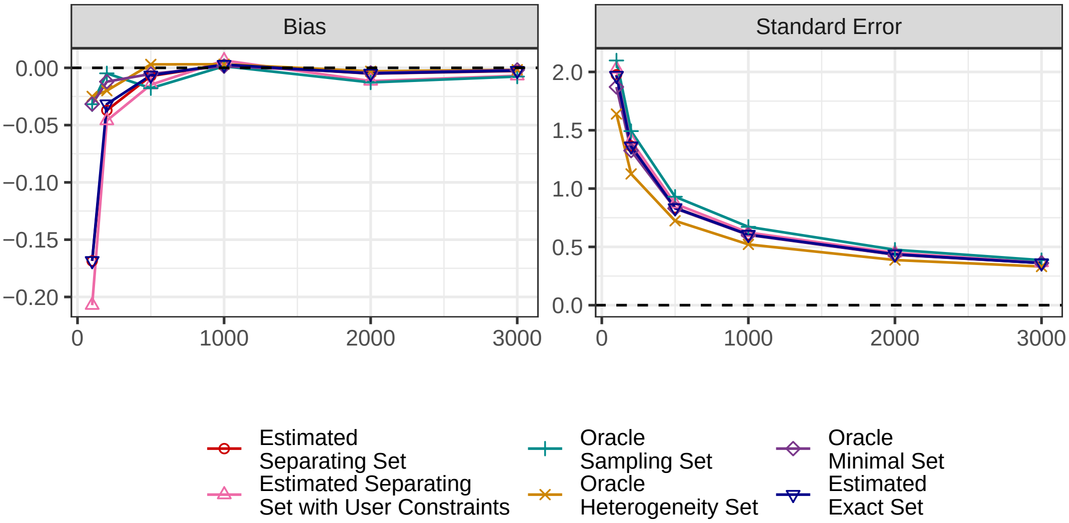

Figure SM-3: Simulation Results. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. More importantly, estimators based on the estimated separating sets (red) and estimated separating sets with user constraints (pink) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple).

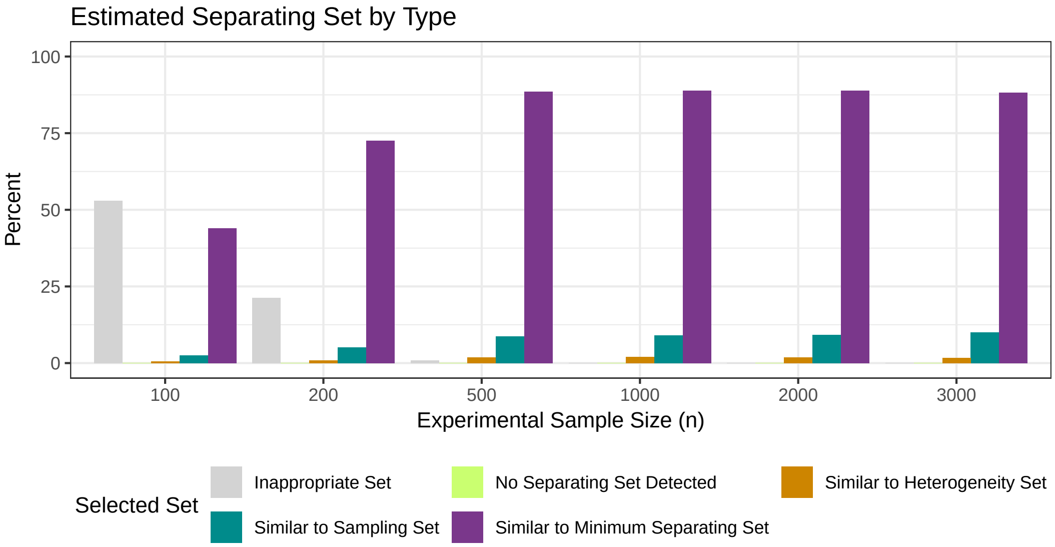

Figure SM-4: Types of Estimated Separating Sets. Note: We present the frequency of estimated separating sets by conceptual type. While the algorithm picks an inappropriate set when the sample size is small, as $n$ increases, the most likely set is the minimal separating set.

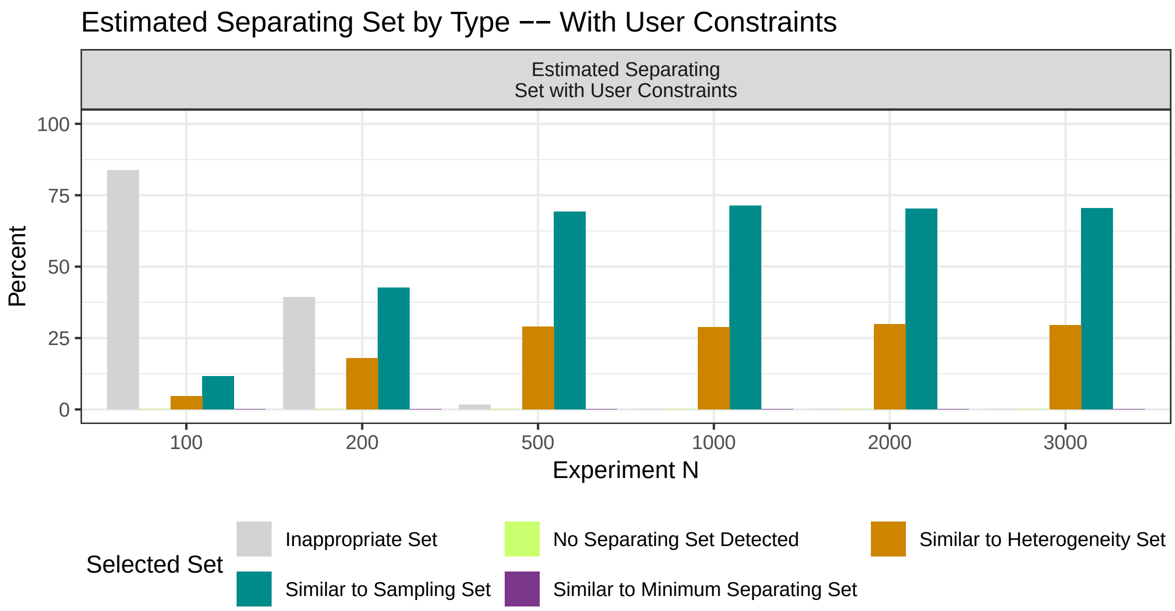

Figure SM-5: Type of Estimated Marginal Separating Set with User Constraints. Note: We present the frequency of estimated separating sets by conceptual type. With user constraints, the algorithm selects each of the other types of separating sets more frequently.

sets are selected drops off rapidly with sample size. In the supplementary material, we show that, when incorporating user constraints that make adjustment by the minimum separating set infeasible, the algorithm selects sets similar to the sampling and heterogeneity sets with higher frequency.

### SM-5.3 Additional Simulation Results

In the previous subsection, we discussed the breakdown of the different types of estimated separating sets in the simulated data generating process. Here we show the breakdown of types of estimated separating sets when incorporating user constraints in Figure SM-5. In this case, $X1$, the alternative separating set, cannot be measured in the target population, we see that the algorithm selects the sampling and heterogeneity sets with higher frequency.

Figure SM-6 presents 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 Figure SM-6). 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.

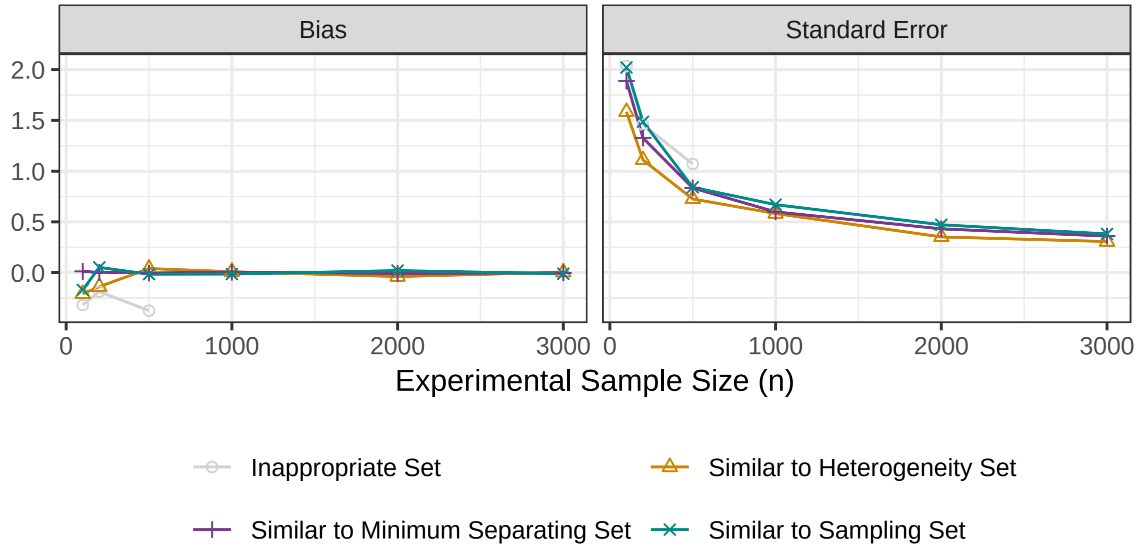

Figure SM-6: Simulation Results for Estimated Separating Set by Type. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. Estimated sets are categorized by type: similar to oracle sampling set (green) and the oracle minimum separating set (purple) and oracle heterogeneity set (orange).

Finally, we present the simulation results for two alternative estimators in Figure SM-7, a calibration estimator and a linear regression projection. The calibration estimator matches population means for the estimated separating set using a maximum entropy (raking) algorithm (Hartman *et al.*, 2015). The linear projection estimator estimates a fully interacted linear regression model using the estimated separating set, and projects the model on the target population.

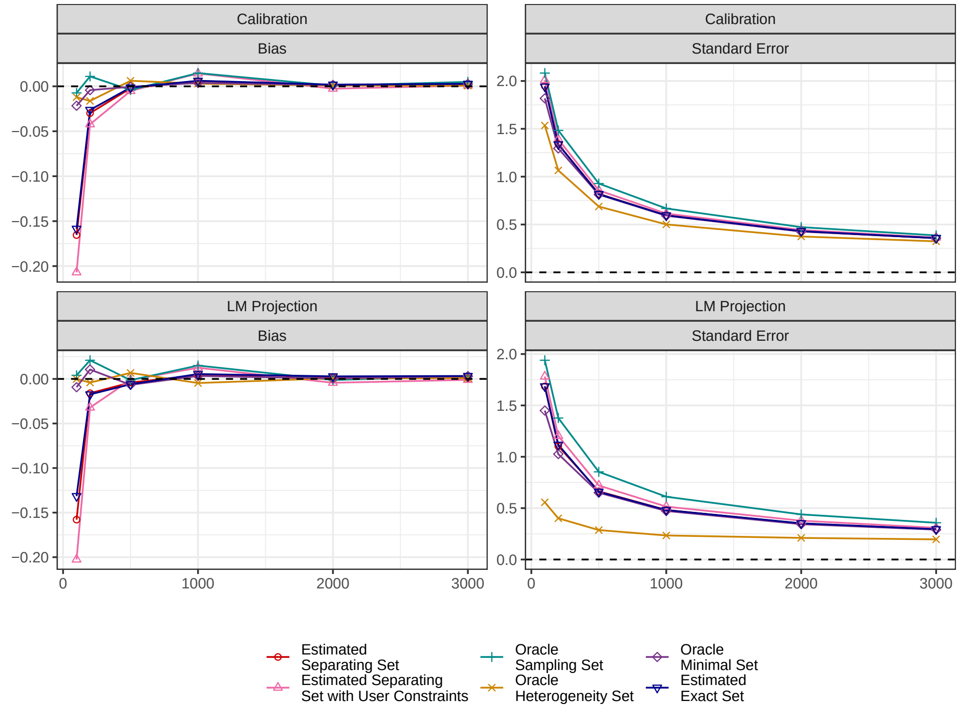

Figure SM-7: Simulation Results for Alternative Estimators. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. More importantly, estimators based on the estimated separating sets (red) and estimated separating set with user constraints (pink) 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.

## SM-6 R Function to Estimate Separating Sets

```
# ################################
# Estimating the separating set
# ################################
# X.data: all pre-treatment covariates in the experimental data
# X.type: types of each covariate. "g" for continous variables, and "c" for categorical variables.
# X.level: the number of levels in each covariates. For continous variables, set it to 1.
# Y: outcome variable in the experimental data
# Treat: treatment variable in the experimental data
# XS: names of the sampling set
# XH: names of the heterogeneity set
# XU: names of variables unmeasurable in the target population
# type: when "Y", we estimate the marginal separating set. when "XH", we estimate the exact separating set.
# print_graph: whether we print the estimated Markov Random Fields

library(igraph); library(qgraph); library(lpSolve); library(mgm); library(Hmisc)

Separating <- function(X.data, X.type, X.level,
                       Y, Treat, XS, XH = NULL, XU=NULL, type = "Y",
                       print_graph = FALSE) {

  ## Setup
  n.var <- ncol(X.data)
  if(type == "Y"){
    if(missing(X.type) == TRUE){
      type.sim  <- rep("g", n.var + 2)
      level.sim <- rep(1, n.var + 2)
    }else{
      type.sim <- c(X.type, rep("g", 2))
      level.sim <- c(X.level, rep(1, 2))
    }
    X.data.g <- cbind(X.data, Y, Treat)
    name.label <- c(colnames(X.data), "Y")
  }else if(type == "XH"){
    if(missing(X.type) == TRUE){
      type.sim  <- rep("g", n.var + 1)
      level.sim <- rep(1, n.var + 1)
    }else{
      type.sim <- c(X.type, rep("g", 1))
      level.sim <- c(X.level, rep(1, 1))
    }
    X.data.g <- cbind(X.data, Treat)
    name.label <- colnames(X.data)
  }
```

```
## ###########################################
## Step 1: Estimate the Markov Random Graph
## ###########################################
fit.sim <- mgm(
  data = X.data.g,
  type = type.sim,
  level = level.sim,
  threshold = "LW",
  k = 2,
  verbatim = TRUE,
  signInfo = FALSE,
  lambdaSel = "EBIC"
)


## Remove T from the Graph
treat_ind <- which(colnames(X.data.g) == "Treat")
Ad <- as.matrix(fit.sim$pairwise$wadj > 0)
Ad <- Ad[-treat_ind,-treat_ind]
Ad.w <- fit.sim$pairwise$wadj
Ad.w <- Ad.w[-treat_ind, -treat_ind]
edge.col <- fit.sim$pairwise$edgecolor
edge.col <- edge.col[-treat_ind,-treat_ind]
graph.u <- graph_from_adjacency_matrix(Ad)

## Show the graph
if(print_graph) qgraph(
  Ad.w,
  edge.color = edge.col,
  layout = ’spring’,
  labels = name.label
)

## #################################################################
## Step 2: Estimate the Separating Set based on an estimated MRF
## #################################################################
if (type == "Y") {
  base <- rep(0, (n.var + 1))
  XS.ind <- which(is.element(colnames(X.data), XS))
  path.cons <- matrix(NA, nrow = 0, ncol = (n.var + 1))
  ## Enumerate all path
  for (w in 1:length(XS)) {
    ind.path.mat <- do.call("rbind",
                            lapply(all_simple_paths(graph.u, (n.var + 1), XS.ind[w]),
                                   FUN=function(x) ind.path(x, base)))
    path.cons <- rbind(path.cons, ind.path.mat)
  }
```

```
}else if (type == "XH") {
  base <- rep(0, n.var)
  XJ  <- intersect(XS, XH)
  all.pair <- expand.grid(XH, XS)
  path.cons <- matrix(NA, nrow = 0, ncol = n.var)
  ## Enumerate all path
  for (w in 1:nrow(all.pair)) {
    ind_1 <- which(colnames(X.data.g) == all.pair[w, 1])
    ind_2 <- which(colnames(X.data.g) == all.pair[w, 2])
    ind.path.mat <-
      do.call("rbind",
              lapply(
                all_simple_paths(graph.u, ind_1, ind_2),
                FUN=function(x) ind.path(x, base)))
    path.cons <- rbind(path.cons, ind.path.mat)
  }
}

if(dim(path.cons)[1] == 0) {
  solution <- NULL
  status <- 0
}else{

  ## Removing Y and XU from the separating set
  if (length(XU) == 0) {
    if(type == "Y"){
      path.cons2 <- rbind(path.cons,
                          c(rep(0, n.var), 1))
      f.dir <- c(rep(">=", nrow(path.cons)), "=")
      f.rhs <- c(rep(1, nrow(path.cons)), 0)
    }else if(type == "XH"){
      path.cons2 <- path.cons
      f.dir <- rep(">=", nrow(path.cons))
      f.rhs <- rep(1, nrow(path.cons))
    }
  } else{
    XU.ind <- which(is.element(colnames(X.data), XU))
    path.cons2.u <- matrix(0, nrow = length(XU.ind), ncol = n.var)
    for (i in 1:nrow(path.cons2.u)) {
      path.cons2.u[i, XU.ind[i]] <- 1
    }
    if(type == "Y"){
      path.cons2.u2 <- cbind(path.cons2.u, 0)
      path.cons2 <- rbind(path.cons,
                          c(rep(0, n.var), 1),
                          path.cons2.u2)
      f.dir <- c(rep(">=", nrow(path.cons)),
                 rep("=", (nrow(path.cons2.u2) + 1)))
      f.rhs <- c(rep(1, nrow(path.cons)),
                 rep(0, (nrow(path.cons2.u2) + 1)))
    }else if(type == "XH"){
      path.cons2.u2 <- path.cons2.u
      path.cons2 <- rbind(path.cons,
                          path.cons2.u2)
      f.dir <- c(rep(">=", nrow(path.cons)),
                 rep("=", nrow(path.cons2.u2)))
      f.rhs <- c(rep(1, nrow(path.cons)),
                 rep(0, nrow(path.cons2.u2)))
    }
  }

  if(type == "Y"){f.obj <- c(rep(1, n.var), 0)}
  else if(type == "XH"){f.obj <- rep(1, n.var)}
  f.con <- path.cons2

  num.solutions <- max.solutions.calculate <- 1
  sp.out <- lp("min", f.obj, f.con, f.dir, f.rhs, all.bin = TRUE, num.bin.solns = max.solutions.calculate)
  if(sp.out$status == 0) {
    if(max.solutions.calculate > 1) {
      solution <- sp.out$solution[1:(length(f.obj)*max.solutions.calculate)]
      solution <- split(solution, sort(1:length(solution) %% sp.out$num.bin.solns))
      if(num.solutions == 1) {
        solution <- sample(solution, num.solutions)
      }
      if(length(solution) == 1) {
        solution <- as.vector(unlist(solution))
      }
    } else {
      solution <- sp.out$solution
    }
  }
  status <- sp.out$status
}

## Final Adjustment
if (status == 0) {
  if(is.null(solution)==TRUE) {
    ## the empty set is enough for generalizability
    solution.name <- NULL
  }else{
    if(type == "Y"){
      solution.ind <-  which(solution[-length(solution)] == 1)
      XJ <- NULL
    }else if (type == "XH") {
        solution.ind <-  which(solution == 1)
      }
      solution.name <- colnames(X.data)[solution.ind]
      if(type == "XH" & length(XJ)!=0){ solution.name <- union(solution.name, XJ)}
    }
  }else if (status==2){
    cat("\nNo Feasible Solution.\n")
    solution.name <- "No Feasible Solution."
  }

  if(print_graph==TRUE){cat ("\n"); cat(solution.name)}

  return(solution.name)
}

# Auxiliary function
ind.path <- function(x, base) {
  base[x] <- 1
  return(base)
}
```

# References

Dahabreh, I. J., Robertson, S. E., Tchetgen Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing Causal Inferences From Individuals In Randomized Trials to All Trial-Eligible Individuals. *Biometrics*, **75**(2), 685–694.

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.

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

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**(427), 846–866.