Local Discovery by Partitioning:

Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs

Jacqueline Maasch | Department of Computer Science

maasch@cs.cornell.edu

16 November 2023

CORNELL TECH

Weill Cornell Medicine

Local Discovery by Partitioning: Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs

Jacqueline R. M. A. Maasch
Cornell Tech

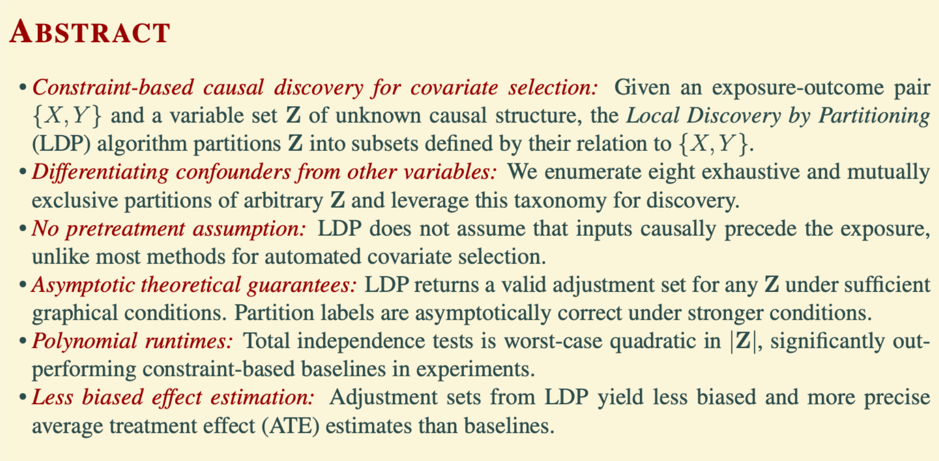
Weishen Pan
Weill Cornell Medicine

Shantanu Gupta
Carnegie Mellon University

Volodymyr Kuleshov
Cornell Tech

Kyra Gan
Cornell Tech

Fei Wang
Weill Cornell Medicine



Abstract

- *Constraint-based causal discovery for covariate selection:* Given an exposure-outcome pair $\{X, Y\}$ and a variable set $\mathbf{Z}$ of unknown causal structure, the *Local Discovery by Partitioning* (LDP) algorithm partitions $\mathbf{Z}$ into subsets defined by their relation to $\{X, Y\}$.
- *Differentiating confounders from other variables:* We enumerate eight exhaustive and mutually exclusive partitions of arbitrary $\mathbf{Z}$ and leverage this taxonomy for discovery.
- *No pretreatment assumption:* LDP does not assume that inputs causally precede the exposure, unlike most methods for automated covariate selection.
- *Asymptotic theoretical guarantees:* LDP returns a valid adjustment set for any $\mathbf{Z}$ under sufficient graphical conditions. Partition labels are asymptotically correct under stronger conditions.
- *Polynomial runtimes:* Total independence tests is worst-case quadratic in $|\mathbf{Z}|$, significantly outperforming constraint-based baselines in experiments.
- *Less biased effect estimation:* Adjustment sets from LDP yield less biased and more precise average treatment effect (ATE) estimates than baselines.

Table of Contents

1 Background

- ▶ Background
- ▶ Preliminaries
- ▶ Partitions of $\mathbf{Z}$
- ▶ Local Discovery by Partitioning (LDP)
- ▶ Empirical results
- ▶ References

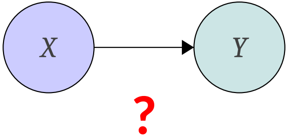
Does X cause Y?

Say you care about two random variables, X and Y.
You know they are statistically dependent, but you don't know why.

$$X \not\perp\!\!\!\perp Y$$

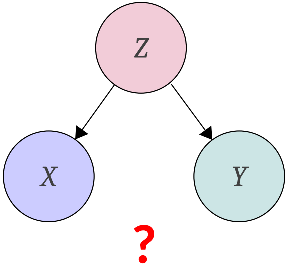
Does X cause Y?

You know from experience that Y could not cause X.
But does X cause Y?
Or is there another explanation?



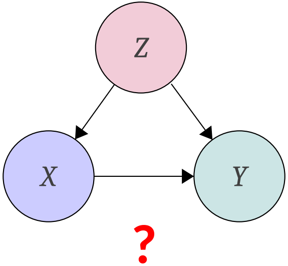
Does X cause Y?

Perhaps a third variable causes both instead.
Here, Z is a *confounder* for X and Y.



Does X cause Y?

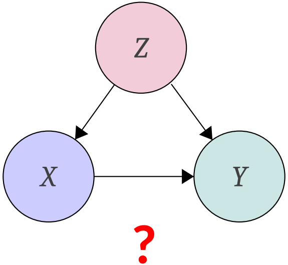
Or perhaps both are true.



Does X cause Y?

1 Background

And if X does cause Y, how strong is the effect?



Covariate adjustment for causal effect estimation

- To obtain an *unbiased* estimate of the causal effect of X on Y, we need to adjust for all *confounders* for $\{X, Y\}$.
- How do we select the variables to adjust for?
- Covariate selection is a central task in the design of observational studies [1].
- The primary goal of covariate selection is to obtain a *valid adjustment set* for an exposure-outcome pair that eliminates *confounding bias* [2].
- Confounding bias distorts the observed relationship between the exposure and outcome, leading to incorrect effect measures *even under infinite data* [3].

Why not adjust for *everything*?

- A naive approach is to adjust for all measured variables.
- However, it is established that multiple variable types can *induce bias* when retained for adjustment [4, 5].
    1. Colliders induce selection bias [6–8].
    2. Mediators bias total effects by controlling for indirect effects [9].
    3. Instruments can amplify existing bias or introduce new bias in some settings [10].
- Further, unnecessary adjustment [5] may increase the variance of causal effect estimates or undermine model fitting due to the curse of dimensionality [11].

Causal discovery for automated covariate selection

In this paper, we address the following question:

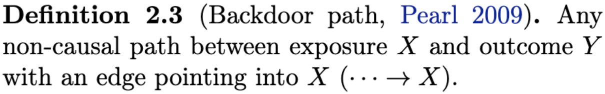
In the absence of prior knowledge, does there exist a polynomial-time algorithm that can select covariates in a principled, automated, and causality-based manner with theoretical guarantees on correctness?

Table of Contents

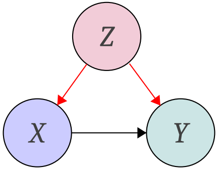
2 Preliminaries

- ▶ Background
- ▶ Preliminaries
- ▶ Partitions of $\mathbf{Z}$
- ▶ Local Discovery by Partitioning (LDP)
- ▶ Empirical results
- ▶ References

Non-causal association flows along open backdoor paths

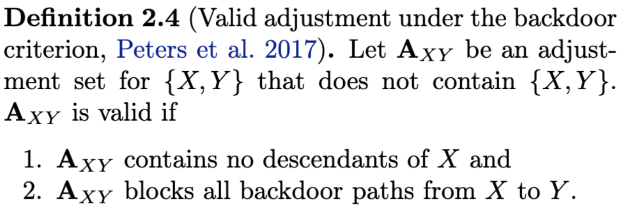


Definition 2.3 (Backdoor path, Pearl 2009)**.** Any non-causal path between exposure X and outcome Y with an edge pointing into X $(\cdots \to X)$.



Valid adjustment under the backdoor criterion

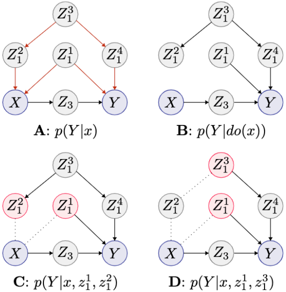
2 Preliminaries



Definition 2.4 (Valid adjustment under the backdoor criterion, Peters et al. 2017)**.** Let $\mathbf{A}_{XY}$ be an adjustment set for $\{X, Y\}$ that does not contain $\{X, Y\}$. $\mathbf{A}_{XY}$ is valid if

1. $\mathbf{A}_{XY}$ contains no descendants of X and
2. $\mathbf{A}_{XY}$ blocks all backdoor paths from X to Y.

Adjustment blocks backdoor paths



A: $p(Y|x)$

B: $p(Y|do(x))$

C: $p(Y|x, z_1^1, z_1^2)$

D: $p(Y|x, z_1^1, z_1^3)$

Figure A.1: Valid adjustment sets. Here, the effect of exposure X on outcome Y is mediated by Z_3. Let $\mathbf{Z}_1 = \{Z_1^1, Z_1^2, Z_1^3, Z_1^4\}$. (**A**) The conditional distribution $p(Y|x)$ fails to isolate the causal association between X and Y due to the open backdoor paths through $\mathbf{Z}_1$, pictured as red arrows. (**B**) We can isolate the causal association between X and Y by intervening on X such that edges $Z_1^2 \to X$ and $Z_1^1 \to X$ are removed. This blocks the non-causal association flowing through these backdoor paths. (**C**) We can identify the interventional distribution $p(Y|do(x))$ via a statistical quantity by conditioning on valid adjustment set $\{Z_1^1, Z_1^2\}$ (highlighted in red), which also blocks the flow of non-causal association. (**D**) Valid adjustment sets are often non-unique. An alternative valid adjustment set for this structure would be $\{Z_1^1, Z_1^3\}$, and still others exist. Figure adapted from Neal (2020).

Assuming pretreatment oversimplifies the problem

Most existing methods for automated covariate selection *assume that inputs causally precede the exposure* [12–16].

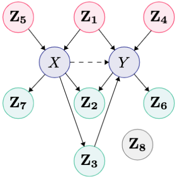


Figure A.2: Pretreatment variables (red) versus post-treatment variables (green). $\mathbf{Z}_1$ (confounders), $\mathbf{Z}_4$, and $\mathbf{Z}_5$ (instruments) are pretreatment variables, which causally precede exposure X. $\mathbf{Z}_2$ (colliders), $\mathbf{Z}_3$ (mediators), $\mathbf{Z}_6$, and $\mathbf{Z}_7$ are post-treatment variables, with X as their causal ancestor. $\mathbf{Z}_8$ is neither pre- nor post-treatment, as it is causally unrelated to X.

This assumption *requires prior knowledge* and *overly simplifies* this task, so we avoid it.

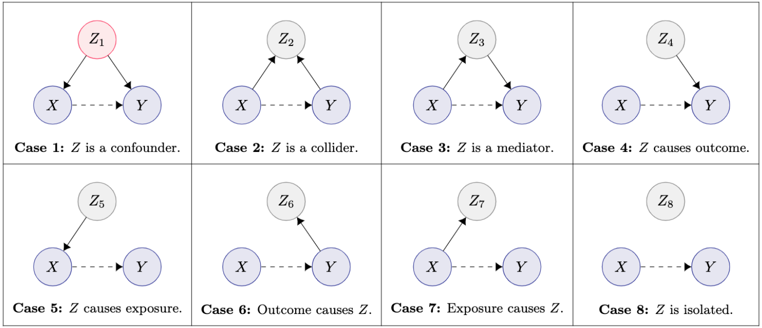
Table of Contents

3 Partitions of $\mathbf{Z}$

- Background
- Preliminaries
- Partitions of $\mathbf{Z}$
- Local Discovery by Partitioning (LDP)
- Empirical results
- References

Triple DAGs: cause, effect, or neither?

3 Partitions of $\mathbf{Z}$



Case 1: Z is a confounder.

Case 2: Z is a collider.

Case 3: Z is a mediator.

Case 4: Z causes outcome.

Case 5: Z causes exposure.

Case 6: Outcome causes Z.

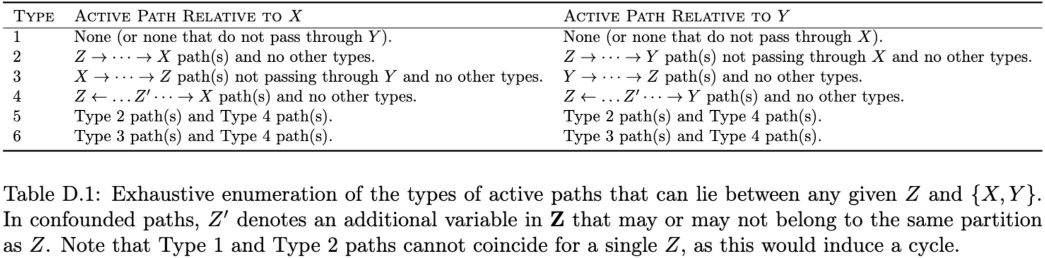
Case 7: Exposure causes Z.

Case 8: Z is isolated.

Table A.1: All potential acyclic triple subgraphs that can be induced by X, Y, and a single Z when paths are restricted to a length of 1. The dashed arrow from exposure X to outcome Y indicates that the strength of this relation is unknown. While the effect of X on Y might be null, it is known that $X \not\perp\!\!\!\perp Y$ and that Y does not cause X. The partition taxonomy proposed in this work (Table 1) generalizes these cases to more complex structures. In the more complex setting, edges represent both direct adjacencies and indirect active paths. Absence of a directed edge therefore indicates either an inactive path or no path at all.

Generalizing to indirect active paths

3 Partitions of $\mathbf{Z}$

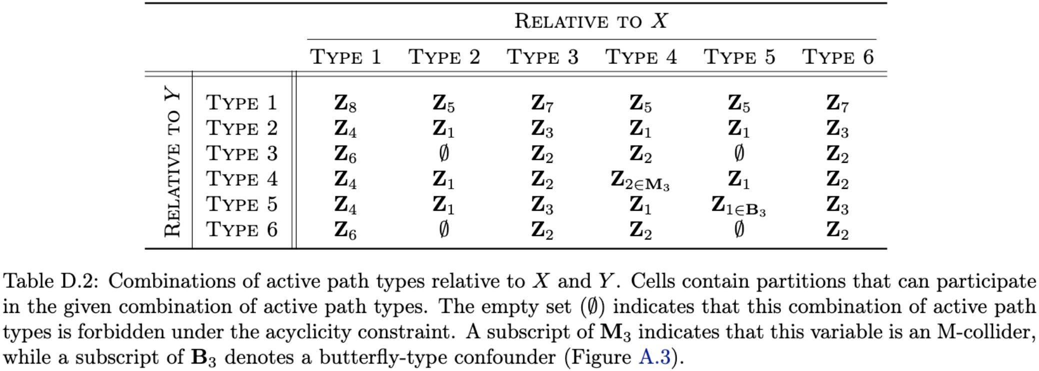


Type	Active Path Relative to X	Active Path Relative to Y
1	None (or none that do not pass through Y).	None (or none that do not pass through X).
2	$Z \to \cdots \to X$ path(s) and no other types.	$Z \to \cdots \to Y$ path(s) not passing through X and no other types.
3	$X \to \cdots \to Z$ path(s) not passing through Y and no other types.	$Y \to \cdots \to Z$ path(s) and no other types.
4	$Z \leftarrow \ldots Z' \cdots \to X$ path(s) and no other types.	$Z \leftarrow \ldots Z' \cdots \to Y$ path(s) and no other types.
5	Type 2 path(s) and Type 4 path(s).	Type 2 path(s) and Type 4 path(s).
6	Type 3 path(s) and Type 4 path(s).	Type 3 path(s) and Type 4 path(s).

Table D.1: Exhaustive enumeration of the types of active paths that can lie between any given Z and $\{X, Y\}$. In confounded paths, Z' denotes an additional variable in $\mathbf{Z}$ that may or may not belong to the same partition as Z. Note that Type 1 and Type 2 paths cannot coincide for a single Z, as this would induce a cycle.

All possible path type combinations

3 Partitions of $\mathbf{Z}$

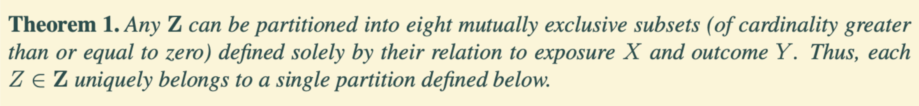


		Relative to X					
		Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
Relative to Y	Type 1	$\mathbf{Z}_8$	$\mathbf{Z}_5$	$\mathbf{Z}_7$	$\mathbf{Z}_5$	$\mathbf{Z}_5$	$\mathbf{Z}_7$
	Type 2	$\mathbf{Z}_4$	$\mathbf{Z}_1$	$\mathbf{Z}_3$	$\mathbf{Z}_1$	$\mathbf{Z}_1$	$\mathbf{Z}_3$
	Type 3	$\mathbf{Z}_6$	$\emptyset$	$\mathbf{Z}_2$	$\mathbf{Z}_2$	$\emptyset$	$\mathbf{Z}_2$
	Type 4	$\mathbf{Z}_4$	$\mathbf{Z}_1$	$\mathbf{Z}_2$	$\mathbf{Z}_{2 \in \mathbf{M}_3}$	$\mathbf{Z}_1$	$\mathbf{Z}_2$
	Type 5	$\mathbf{Z}_4$	$\mathbf{Z}_1$	$\mathbf{Z}_3$	$\mathbf{Z}_1$	$\mathbf{Z}_{1 \in \mathbf{B}_3}$	$\mathbf{Z}_3$
	Type 6	$\mathbf{Z}_6$	$\emptyset$	$\mathbf{Z}_2$	$\mathbf{Z}_2$	$\emptyset$	$\mathbf{Z}_2$

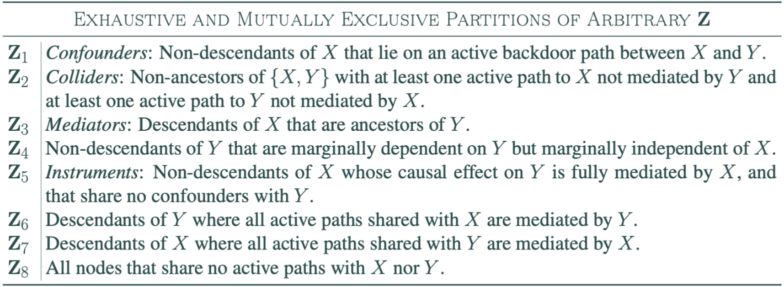
Table D.2: Combinations of active path types relative to X and Y. Cells contain partitions that can participate in the given combination of active path types. The empty set ($\emptyset$) indicates that this combination of active path types is forbidden under the acyclicity constraint. A subscript of $\mathbf{M}_3$ indicates that this variable is an M-collider, while a subscript of $\mathbf{B}_3$ denotes a butterfly-type confounder (Figure A.3).

A taxonomy for arbitrary Z

3 Partitions of $\mathbf{Z}$



Theorem 1. *Any* $\mathbf{Z}$ *can be partitioned into eight mutually exclusive subsets (of cardinality greater than or equal to zero) defined solely by their relation to exposure* X *and outcome* Y*. Thus, each* $Z \in \mathbf{Z}$ *uniquely belongs to a single partition defined below.*



Exhaustive and Mutually Exclusive Partitions of Arbitrary $\mathbf{Z}$

$\mathbf{Z}_1$	*Confounders*: Non-descendants of X that lie on an active backdoor path between X and Y.
$\mathbf{Z}_2$	*Colliders*: Non-ancestors of $\{X, Y\}$ with at least one active path to X not mediated by Y and at least one active path to Y not mediated by X.
$\mathbf{Z}_3$	*Mediators*: Descendants of X that are ancestors of Y.
$\mathbf{Z}_4$	Non-descendants of Y that are marginally dependent on Y but marginally independent of X.
$\mathbf{Z}_5$	*Instruments*: Non-descendants of X whose causal effect on Y is fully mediated by X, and that share no confounders with Y.
$\mathbf{Z}_6$	Descendants of Y where all active paths shared with X are mediated by Y.
$\mathbf{Z}_7$	Descendants of X where all active paths shared with Y are mediated by X.
$\mathbf{Z}_8$	All nodes that share no active paths with X nor Y.

Reduction to the 10-node DAG

3 Partitions of $\mathbf{Z}$

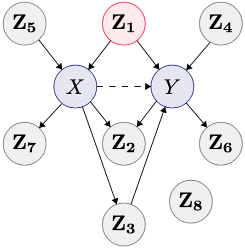


Figure 1: The partitions of $\mathbf{Z}$ (Table 1) reduce to a 10-node DAG surrounding $\{X, Y\}$ where nodes represent partition sets, arrows signify both direct adjacencies and indirect active paths (one or more), and inter-covariate paths are abstracted away. The dashed edge between X and Y indicates that the strength of this relation is unknown, and may be null. Conditioning on $\mathbf{Z}_1$ in red blocks all backdoor paths for $\{X, Y\}$.

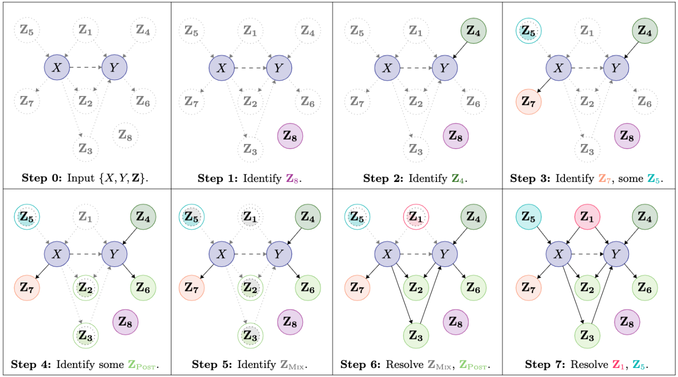
Table of Contents

4 Local Discovery by Partitioning (LDP)

- Background
- Preliminaries
- Partitions of $\mathbf{Z}$
- Local Discovery by Partitioning (LDP)
- Empirical results
- References

Visual intuition

4 Local Discovery by Partitioning (LDP)



Step 0: Input $\{X, Y, \mathbf{Z}\}$.

Step 1: Identify $\mathbf{Z}_8$.

Step 2: Identify $\mathbf{Z}_4$.

Step 3: Identify $\mathbf{Z}_7$, some $\mathbf{Z}_5$.

Step 4: Identify some $\mathbf{Z}_{\text{Post}}$.

Step 5: Identify $\mathbf{Z}_{\text{Mix}}$.

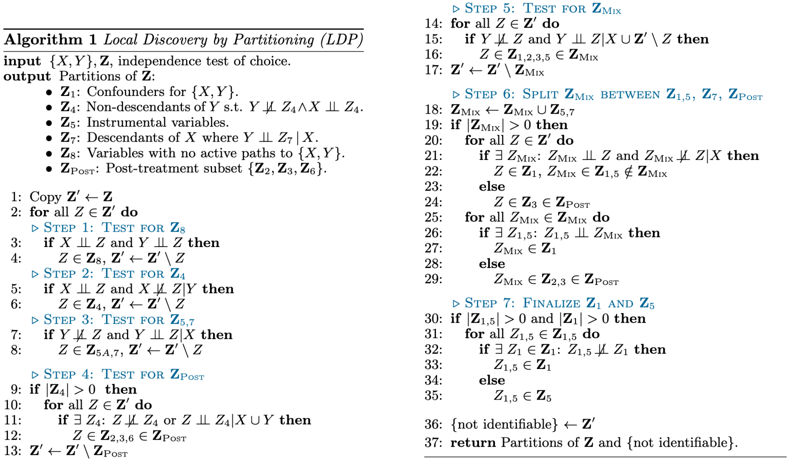
Step 6: Resolve $\mathbf{Z}_{\text{Mix}}$, $\mathbf{Z}_{\text{Post}}$.

Step 7: Resolve $\mathbf{Z}_1$, $\mathbf{Z}_5$.

Table D.3: Schematic of Algorithm 1. The exposure-outcome pair $\{X, Y\}$ serves as a nucleus around which LDP assembles a partial causal graph. Each step reveals additional information about the partitions of $\mathbf{Z}$. Nodes that are fully colored are fully discovered by Algorithm 1. Partial coloring denotes partial knowledge, and no coloring denotes no knowledge.

Pseudocode

4 Local Discovery by Partitioning (LDP)



Algorithm 1 *Local Discovery by Partitioning (LDP)*

input $\{X, Y\}, \mathbf{Z}$, independence test of choice.

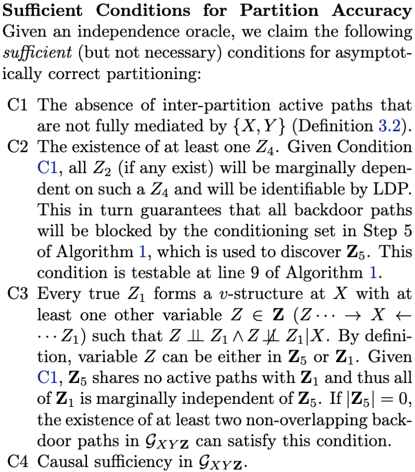
output Partitions of $\mathbf{Z}$:

- $\mathbf{Z}_1$: Confounders for $\{X, Y\}$.
- $\mathbf{Z}_4$: Non-descendants of Y s.t. $Y \not\perp\!\!\!\perp Z_4 \wedge X \perp\!\!\!\perp Z_4$.
- $\mathbf{Z}_5$: Instrumental variables.
- $\mathbf{Z}_7$: Descendants of X where $Y \perp\!\!\!\perp Z_7 \,|\, X$.
- $\mathbf{Z}_8$: Variables with no active paths to $\{X, Y\}$.
- $\mathbf{Z}_{\text{Post}}$: Post-treatment subset $\{\mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_6\}$.

```
Copy $\mathbf{Z}' \leftarrow \mathbf{Z}$
for all $Z \in \mathbf{Z}'$ do
    ▷ Step 1: Test for $\mathbf{Z}_8$
    if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$ then
        $Z \in \mathbf{Z}_8,\ \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$
    ▷ Step 2: Test for $\mathbf{Z}_4$
    if $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z|Y$ then
        $Z \in \mathbf{Z}_4,\ \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$
    ▷ Step 3: Test for $\mathbf{Z}_{5,7}$
    if $Y \not\perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z|X$ then
        $Z \in \mathbf{Z}_{5A,7},\ \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$

    ▷ Step 4: Test for $\mathbf{Z}_{\text{Post}}$
if $|\mathbf{Z}_4| > 0$ then
    for all $Z \in \mathbf{Z}'$ do
        if $\exists\, Z_4$: $Z \not\perp\!\!\!\perp Z_4$ or $Z \perp\!\!\!\perp Z_4|X \cup Y$ then
            $Z \in \mathbf{Z}_{2,3,6} \in \mathbf{Z}_{\text{Post}}$
$\mathbf{Z}' \leftarrow \mathbf{Z}' \setminus \mathbf{Z}_{\text{Post}}$
    ▷ Step 5: Test for $\mathbf{Z}_{\text{Mix}}$
for all $Z \in \mathbf{Z}'$ do
    if $Y \not\perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z|X \cup \mathbf{Z}' \setminus Z$ then
        $Z \in \mathbf{Z}_{1,2,3,5} \in \mathbf{Z}_{\text{Mix}}$
$\mathbf{Z}' \leftarrow \mathbf{Z}' \setminus \mathbf{Z}_{\text{Mix}}$

    ▷ Step 6: Split $\mathbf{Z}_{\text{Mix}}$ between $\mathbf{Z}_{1,5}$, $\mathbf{Z}_7$, $\mathbf{Z}_{\text{Post}}$
$\mathbf{Z}_{\text{Mix}} \leftarrow \mathbf{Z}_{\text{Mix}} \cup \mathbf{Z}_{5,7}$
if $|\mathbf{Z}_{\text{Mix}}| > 0$ then
    for all $Z \in \mathbf{Z}'$ do
        if $\exists\, Z_{\text{Mix}}$: $Z_{\text{Mix}} \perp\!\!\!\perp Z$ and $Z_{\text{Mix}} \not\perp\!\!\!\perp Z|X$ then
            $Z \in \mathbf{Z}_1,\ Z_{\text{Mix}} \in \mathbf{Z}_{1,5} \notin \mathbf{Z}_{\text{Mix}}$
        else
            $Z \in \mathbf{Z}_3 \in \mathbf{Z}_{\text{Post}}$
    for all $Z_{\text{Mix}} \in \mathbf{Z}_{\text{Mix}}$ do
        if $\exists\, Z_{1,5}$: $Z_{1,5} \perp\!\!\!\perp Z_{\text{Mix}}$ then
            $Z_{\text{Mix}} \in \mathbf{Z}_1$
        else
            $Z_{\text{Mix}} \in \mathbf{Z}_{2,3} \in \mathbf{Z}_{\text{Post}}$

    ▷ Step 7: Finalize $\mathbf{Z}_1$ and $\mathbf{Z}_5$
if $|\mathbf{Z}_{1,5}| > 0$ and $|\mathbf{Z}_1| > 0$ then
    for all $Z_{1,5} \in \mathbf{Z}_{1,5}$ do
        if $\exists\, Z_1 \in \mathbf{Z}_1$: $Z_{1,5} \not\perp\!\!\!\perp Z_1$ then
            $Z_{1,5} \in \mathbf{Z}_1$
        else
            $Z_{1,5} \in \mathbf{Z}_5$

{not identifiable} $\leftarrow \mathbf{Z}'$
return Partitions of $\mathbf{Z}$ and {not identifiable}.
```

Sufficient conditions for identifiability

4 Local Discovery by Partitioning (LDP)



Sufficient Conditions for Partition Accuracy Given an independence oracle, we claim the following *sufficient* (but not necessary) conditions for asymptotically correct partitioning:

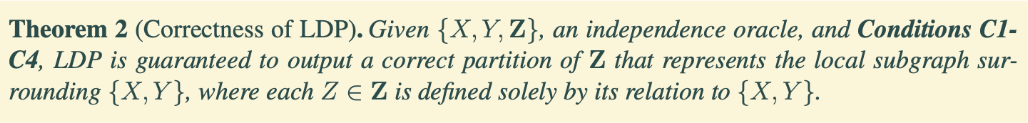
C1 The absence of inter-partition active paths that are not fully mediated by $\{X, Y\}$ (Definition 3.2).

C2 The existence of at least one Z_4. Given Condition C1, all Z_2 (if any exist) will be marginally dependent on such a Z_4 and will be identifiable by LDP. This in turn guarantees that all backdoor paths will be blocked by the conditioning set in Step 5 of Algorithm 1, which is used to discover $\mathbf{Z}_5$. This condition is testable at line 9 of Algorithm 1.

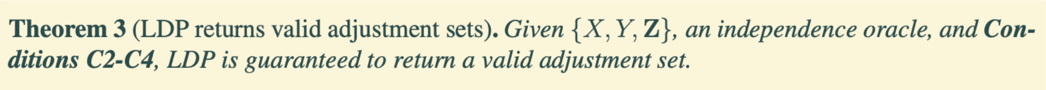
C3 Every true Z_1 forms a *v*-structure at X with at least one other variable $Z \in \mathbf{Z}$ ($Z \cdots \to X \leftarrow \cdots Z_1$) such that $Z \perp\!\!\!\perp Z_1 \land Z \not\perp\!\!\!\perp Z_1 | X$. By definition, variable Z can be either in $\mathbf{Z}_5$ or $\mathbf{Z}_1$. Given C1, $\mathbf{Z}_5$ shares no active paths with $\mathbf{Z}_1$ and thus all of $\mathbf{Z}_1$ is marginally independent of $\mathbf{Z}_5$. If $|\mathbf{Z}_5| = 0$, the existence of at least two non-overlapping backdoor paths in $\mathcal{G}_{XY\mathbf{Z}}$ can satisfy this condition.

C4 Causal sufficiency in $\mathcal{G}_{XY\mathbf{Z}}$.

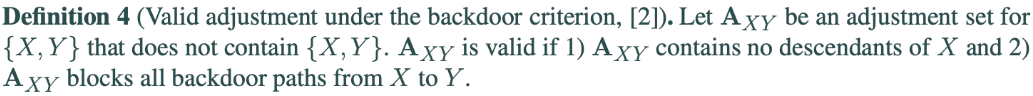
Guarantees on correctness



Theorem 2 (Correctness of LDP). *Given* $\{X, Y, \mathbf{Z}\}$*, an independence oracle, and* ***Conditions C1-C4****, LDP is guaranteed to output a correct partition of* $\mathbf{Z}$ *that represents the local subgraph surrounding* $\{X, Y\}$*, where each* $Z \in \mathbf{Z}$ *is defined solely by its relation to* $\{X, Y\}$*.*



Theorem 3 (LDP returns valid adjustment sets). *Given* $\{X, Y, \mathbf{Z}\}$*, an independence oracle, and* ***Conditions C2-C4****, LDP is guaranteed to return a valid adjustment set.*



Definition 4 (Valid adjustment under the backdoor criterion, [2]). Let $\mathbf{A}_{XY}$ be an adjustment set for $\{X, Y\}$ that does not contain $\{X, Y\}$. $\mathbf{A}_{XY}$ is valid if 1) $\mathbf{A}_{XY}$ contains no descendants of X and 2) $\mathbf{A}_{XY}$ blocks all backdoor paths from X to Y.

Time complexity is quadratic in $|\mathbf{Z}|$

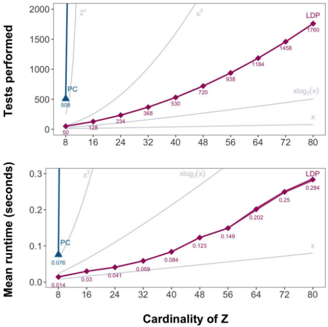


Figure 2: Total tests performed per $\mathbf{Z}$ under an independence oracle (top) and mean runtime over 100 replicates (bottom) as the cardinality of $\mathbf{Z}$ increases, with 95% confidence intervals in shaded regions. Each DAG resembles Figure 1 with equal cardinality per partition ($[1, 10]$).

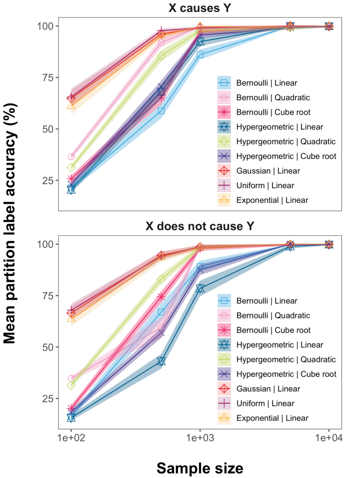
Table of Contents

5 Empirical results

- Background
- Preliminaries
- Partitions of $\mathbf{Z}$
- Local Discovery by Partitioning (LDP)
- Empirical results
- References

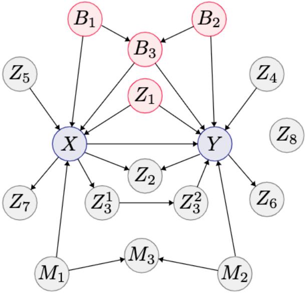
LDP accurately partitions the 10-node DAG

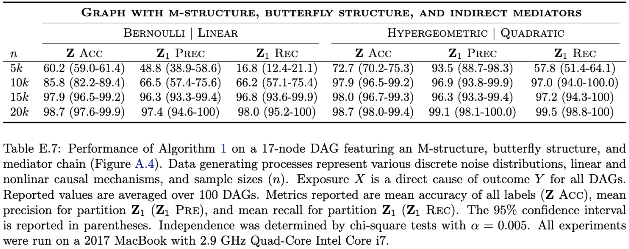
5 Empirical results



LDP accurately partitions more complex DAGs

5 Empirical results



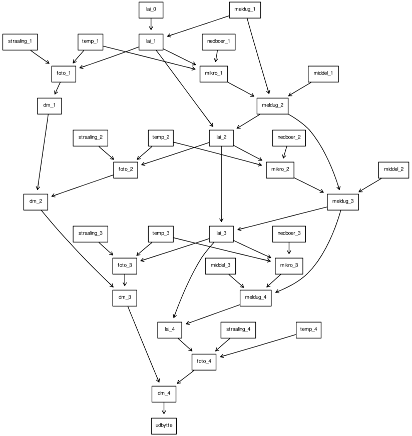


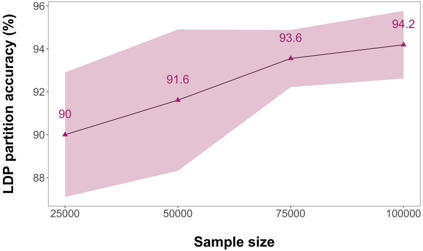
	Graph with M-structure, butterfly structure, and indirect mediators					
	Bernoulli | Linear			Hypergeometric | Quadratic		
n	**Z** Acc	$\mathbf{Z}_1$ Prec	$\mathbf{Z}_1$ Rec	**Z** Acc	$\mathbf{Z}_1$ Prec	$\mathbf{Z}_1$ Rec
$5k$	60.2 (59.0-61.4)	48.8 (38.9-58.6)	16.8 (12.4-21.1)	72.7 (70.2-75.3)	93.5 (88.7-98.3)	57.8 (51.4-64.1)
$10k$	85.8 (82.2-89.4)	66.5 (57.4-75.6)	66.2 (57.1-75.4)	97.9 (96.5-99.2)	96.9 (93.8-99.9)	97.0 (94.0-100.0)
$15k$	97.9 (96.5-99.2)	96.3 (93.3-99.4)	96.8 (93.6-99.9)	98.0 (96.7-99.3)	96.3 (93.3-99.4)	97.2 (94.3-100)
$20k$	98.7 (97.6-99.9)	97.4 (94.6-100)	98.0 (95.2-100)	98.7 (98.0-99.4)	99.1 (98.1-100.0)	99.5 (98.8-100)

Table E.7: Performance of Algorithm 1 on a 17-node DAG featuring an M-structure, butterfly structure, and mediator chain (Figure A.4). Data generating processes represent various discrete noise distributions, linear and nonlinar causal mechanisms, and sample sizes (n). Exposure X is a direct cause of outcome Y for all DAGs. Reported values are averaged over 100 DAGs. Metrics reported are mean accuracy of all labels (**Z** Acc), mean precision for partition $\mathbf{Z}_1$ ($\mathbf{Z}_1$ Pre), and mean recall for partition $\mathbf{Z}_1$ ($\mathbf{Z}_1$ Rec). The 95% confidence interval is reported in parentheses. Independence was determined by chi-square tests with $\alpha = 0.005$. All experiments were run on a 2017 MacBook with 2.9 GHz Quad-Core Intel Core i7.

LDP accurately partitions more complex DAGs

5 Empirical results





The `mildew` benchmark [17] from `bnlearn` [18]. $|\mathbf{Z}| = 31$ for exposure-outcome pair (`mikro_1` $\to$ `meldug_2`).

LDP enables less biased and more precise ATE estimates

5 Empirical results

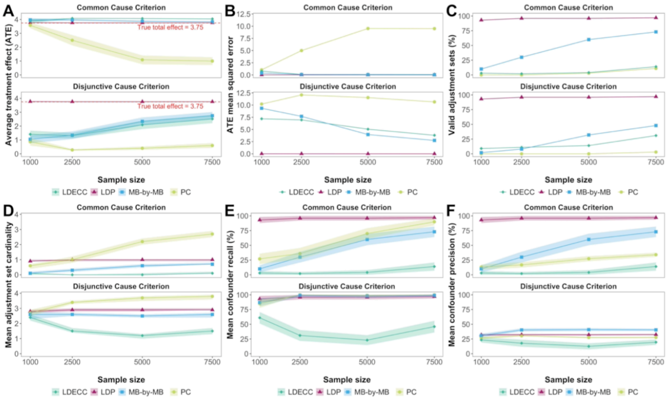


Figure 2: ATE estimation using adjustment sets produced by each baseline for a linear-Gaussian 10-node DAG (Fig. 1). Independence was determined by Fisher-z tests ($\alpha = 0.01$). Results are for 100 replicates per sample size with 95% confidence intervals in shaded regions.

Even if partitions are wrong, adjustment sets are valid

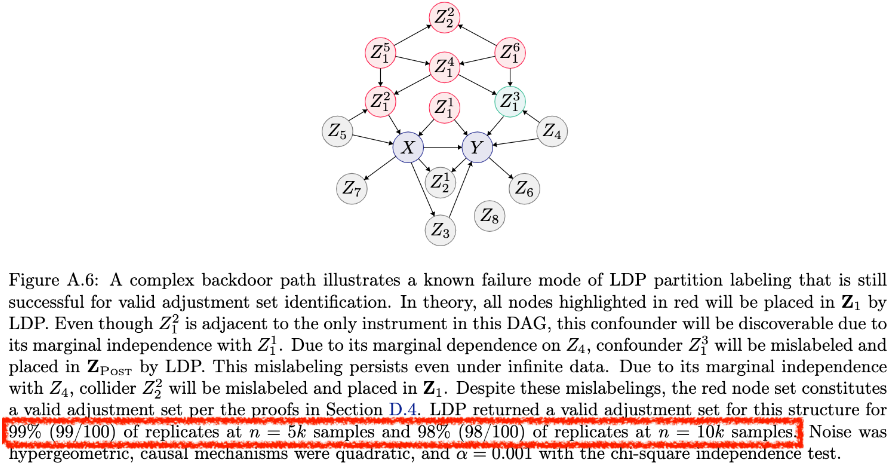


Figure A.6: A complex backdoor path illustrates a known failure mode of LDP partition labeling that is still successful for valid adjustment set identification. In theory, all nodes highlighted in red will be placed in $\mathbf{Z}_1$ by LDP. Even though Z_1^2 is adjacent to the only instrument in this DAG, this confounder will be discoverable due to its marginal independence with Z_1^1. Due to its marginal dependence on Z_4, confounder Z_1^3 will be mislabeled and placed in $\mathbf{Z}_{\text{POST}}$ by LDP. This mislabeling persists even under infinite data. Due to its marginal independence with Z_4, collider Z_2^2 will be mislabeled and placed in $\mathbf{Z}_1$. Despite these mislabelings, the red node set constitutes a valid adjustment set per the proofs in Section D.4. LDP returned a valid adjustment set for this structure for 99% (99/100) of replicates at $n = 5k$ samples and 98% (98/100) of replicates at $n = 10k$ samples. Noise was hypergeometric, causal mechanisms were quadratic, and $\alpha = 0.001$ with the chi-square independence test.

Even if partitions are wrong, adjustment sets are valid

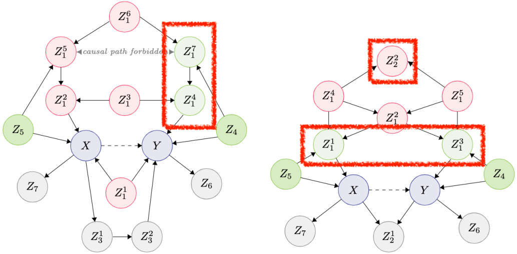


Figure D.1: Two DAGs that exemplify the behavior of LDP for valid adjustment set detection in the presence of inter-partition active paths. All red nodes will be placed in $\mathbf{Z}_1$ by LDP. All confounders for $\{X, Y\}$ that are colored green will be mislabeled due to their marginal dependence on Z_4 or Z_5.

Left: Per Lemma D.20, Z_1^1, Z_1^3 and Z_1^6 will be placed in $\mathbf{Z}_1$. Despite their marginal dependence on the only Z_5 in this structure, Z_1^2 and Z_1^5 will never be placed in $\mathbf{Z}_{\text{Post}}$ due to the presence of Z_1^1, as $Z_1^2 \perp\!\!\!\perp Z_1^1$ and $Z_1^5 \perp\!\!\!\perp Z_1^1$. Together, the confounders highlighted in red ($\{Z_1^1, Z_1^2, Z_1^3, Z_1^5, Z_1^6\}$) constitute a valid adjustment set that blocks all backdoor paths and contains no descendents of X. No causal path of either directionality is permissible between Z_1^5 and Z_1^7 per Proposition D.18. If this path were to contain a confounder analogous to Z_1^3, this would be permissible and this node would be placed in $\mathbf{Z}_1$ by LDP.

Right: This DAG contains a modified butterfly structure, which will be partially retained in $\mathbf{Z}_1$ ($\{Z_1^2, Z_1^4, Z_1^5\}$) while still blocking all backdoor paths. As there is only one Z_5 in this structure and no backdoor path whose members are marginally independent of Z_1^1, this confounder will be mislabeled as $\mathbf{Z}_{\text{Post}}$ at Step 6. This DAG also illustrates a case where a member of $\mathbf{Z}_2$ (Z_2^2) is placed in $\mathbf{Z}_1$. Inclusion of Z_2^2 does not violate the validity of the adjustment set returned by LDP, as this node is not a descendent of X and adjusting for $\{Z_1^2, Z_1^4, Z_1^5\}$ prevents collider bias.

Next steps...

Applying LDP to:

1. Fairness in organ transplantation.
2. Covariate selection for trial emulation.
3. Instrumental variable discovery for Mendelian randomization.

Thank you! Any questions?

maasch@cs.cornell.edu

Table of Contents

6 References

- ▶ Background
- ▶ Preliminaries
- ▶ Partitions of $\mathbb{Z}$
- ▶ Local Discovery by Partitioning (LDP)
- ▶ Empirical results
- ▶ References

References

[1] F. R. Guo et al. *Confounder Selection: Objectives and Approaches*. en. arXiv:2208.13871 [math, stat]. 2022.

[2] J. Witte et al. “Covariate selection strategies for causal inference: Classification and comparison”. en. In: *Biometrical Journal* 61.5 (2019), pp. 1270–1289. DOI: 10.1002/bimj.201700294.

[3] M. A. Hernán et al. *Causal Inference: What If*. en. 2020.

[4] H. Lu et al. “Revisiting Overadjustment Bias”. en. In: *Epidemiology* 32.5 (2021), e22–e23. DOI: 10.1097/EDE.0000000000001377.

[5] E. F. Schisterman et al. “Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies”. en. In: *Epidemiology* 20.4 (2009), pp. 488–495. DOI: 10.1097/EDE.0b013e3181a819a1.

[6] M. A. Hernán et al. “A Structural Approach to Selection Bias”. en. In: *Epidemiology* 15.5 (2004), pp. 615–625. DOI: 10.1097/01.ede.0000135174.63482.43.

[7] F. Elwert et al. “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable”. In: *Annual Review of Sociology* 40.1 (2014), pp. 31–53. DOI: 10.1146/annurev-soc-071913-043455.

[8] M. J. Holmberg et al. “Collider bias”. In: *JAMA Guide to Statistics and Methods* 327.13 (2022).

[9] J. Pearl. “Direct and Indirect Effects”. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. 2001.

[10] J. Pearl. “On a Class of Bias-Amplifying Variables that Endanger Effect Estimates”. en. In: (2012).

References

[11] M. E. Schnitzer et al. “Variable Selection for Confounder Control, Flexible Modeling and Collaborative Targeted Minimum Loss-Based Estimation in Causal Inference”. In: *The International Journal of Biostatistics* 12.1 (2016), pp. 97–115. DOI: 10.1515/ijb-2015-0017.

[12] X. De Luna et al. “Covariate selection for the nonparametric estimation of an average treatment effect”. en. In: *Biometrika* 98.4 (2011), pp. 861–875. DOI: 10.1093/biomet/asr041.

[13] D. Entner et al. “Data-driven covariate selection for nonparametric estimation of causal effects”. en. In: *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2013.

References

[14] S. M. Shortreed et al. “Outcome‐adaptive lasso: Variable selection for causal inference”. en. In: *Biometrics* 73.4 (2017), pp. 1111–1122. DOI: `10.1111/biom.12679`.

[15] Y. Tian et al. “Evaluating large-scale propensity score performance through real-world and synthetic data experiments”. In: *International Journal of Epidemiology* 47.6 (2018), pp. 2005–2014. DOI: `10.1093/ije/dyy120`.

[16] A. Soleymani et al. “Causal Feature Selection via Orthogonal Search”. en. In: *Transactions on Machine Learning Research* (2022). arXiv:2007.02938 [cs, math, stat].

References

[17] A. L. Jensen et al. “MIDAS: An influence diagram for management of mildew in winter wheat”. In: *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*. 1996, pp. 349–356.

[18] M. Scutari. *Learning Bayesian Networks with the bnlearn R Package*. en. arXiv:0908.3817 [stat]. 2010.