Efficient Identification in Linear Structural Causal Models with Instrumental Cutsets

Daniel Kumor[1], Bryant Chen[2], and Elias Bareinboim[3]

[1]Purdue University, [2]Brex Inc, [3]Columbia University

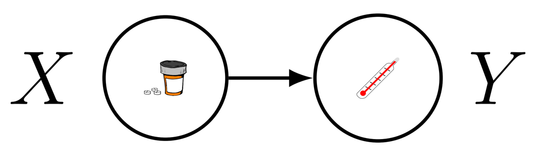
Linear Causal Models and Identification

Problem 1:

There is a medicine X. How to determine whether the medicine can help in treating disease Y?

One can perform randomized clinical trials to assess the effect of X on Y:

$$X = \epsilon_x$$
$$Y = \lambda_{xy} X + \epsilon_y$$



Here, the correlation *is* the causal effect:

$$\sigma_{xy} = \mathrm{E}[XY] = \mathrm{E}[X(\lambda_{xy}X + \epsilon_y)] = \lambda_{xy}\mathrm{E}[XX] + \mathrm{E}[\epsilon_x \epsilon_y] = \lambda_{xy}$$

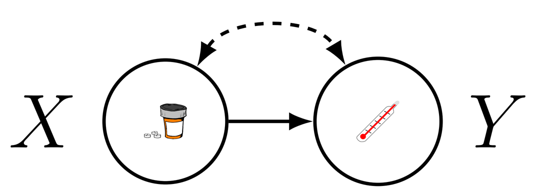
Problem 2:

The medicine has extreme side-effects, so only some patients will agree to try it.

The self-selection of patients into groups adds confounding bias:

$$X = \epsilon_x$$
$$Y = \lambda_{xy} X + \epsilon_y$$

ϵ_x, ϵ_y correlated



Now, the causal effect can no longer be uniquely determined from the data, since we cannot isolate how much of the correlation is due to the causal effect and how much is due to the confounder. In other words, it is *not identifiable*.

Task: Linear Identifiability

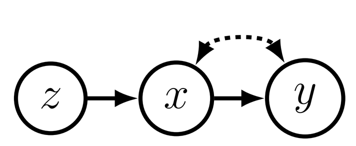
Given a causal graph G, and a target parameter λ_{xy}, is it possible to uniquely determine λ_{xy} when the causal mechanisms are linear?

Approach 1: Instrumental Variables

There exist models that are *not identifiable* without parametric assumptions, such as linearity. When all paths from a variable z to y pass through x, z is called an *instrumental variable* (IV).

$$Z = \epsilon_z$$
$$X = \lambda_{zx} Z + \epsilon_x$$
$$Y = \lambda_{xy} X + \epsilon_y$$

ϵ_x, ϵ_y correlated



Using this IV, we can identify λ_{xy}

$$\sigma_{zy} = \mathrm{E}[ZY] = \mathrm{E}[Z(\lambda_{xy}X + \epsilon_y)] = \lambda_{xy}\mathrm{E}[ZX] + \mathrm{E}[Z\epsilon_y]$$
$$\sigma_{zy} = \lambda_{xy}\sigma_{zx} \quad \Rightarrow \quad \lambda_{xy} = \frac{\sigma_{zy}}{\sigma_{zx}}$$

Several recent identification methods build upon IVs:

1. Conditional Auxiliary Variables (cAV, Chen et al. 2017)

   Allows using a conditioning set to block paths from the instrument to target, and exploits previously solved edges to help in identification. Comes with a polynomial-time algorithm.

2. Trek-Separation IVs (tsIV, Weihs et al. 2017)

   Exploits determinantal constraints in the covariance matrix to block paths from the instrument to target.

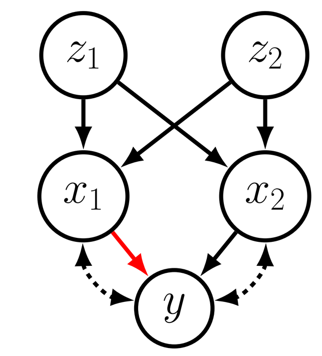
Approach 2: Instrumental Sets

The ideas behind the IV can be extended to *instrumental sets* (IS):

$$Z_1 = \epsilon_{z_1}$$
$$Z_2 = \epsilon_{z_2}$$
$$X_1 = \lambda_{z_1x_1} Z_1 + \lambda_{z_2x_1} Z_2 + \epsilon_{x_1}$$
$$X_2 = \lambda_{z_1x_2} Z_1 + \lambda_{z_2x_2} Z_2 + \epsilon_{x_2}$$
$$Y = \lambda_{x_1y} X_1 + \lambda_{x_2y} X_2 + \epsilon_y$$

$\epsilon_{x_1}, \epsilon_y$ correlated

$\epsilon_{x_2}, \epsilon_y$ correlated



Here, $x_1 \rightarrow y$ has no instruments - both z_1 and z_2 have paths to y through x_2. Nevertheless, we can get a linear system:

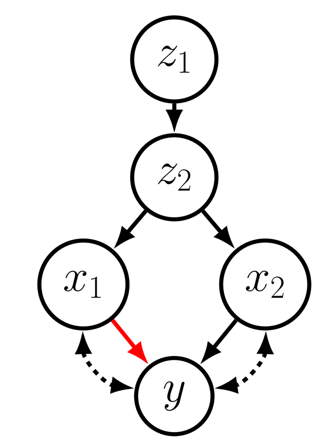
$$\sigma_{z_1y} = \sigma_{z_1x_1}\lambda_{x_1y} + \sigma_{z_1x_2}\lambda_{x_2y}$$
$$\sigma_{z_2y} = \sigma_{z_2x_1}\lambda_{x_1y} + \sigma_{z_2x_2}\lambda_{x_2y}$$

An instrumental set exists for λ_{x_1y} only if $\det\begin{pmatrix}\sigma_{z_1x_1} & \sigma_{z_1x_2} \\ \sigma_{z_2x_1} & \sigma_{z_2x_2}\end{pmatrix} \neq 0$. An example where this requirement is not satisfied is:

$$Z_1 = \epsilon_{z_1}$$
$$Z_2 = \lambda_{z_1z_2} Z_1 + \epsilon_{z_2}$$
$$X_1 = \lambda_{z_2x_1} Z_2 + \epsilon_{x_1}$$
$$X_2 = \lambda_{z_2x_2} Z_2 + \epsilon_{x_2}$$
$$Y = \lambda_{x_1y} X_1 + \lambda_{x_2y} X_2 + \epsilon_y$$

$\epsilon_{x_1}, \epsilon_y$ correlated

$\epsilon_{x_2}, \epsilon_y$ correlated



In this example, the system of equations is degenerate, meaning that $(\sigma_{z_1x_1}, \sigma_{z_1x_2}) = \lambda_{z_1z_2}(\sigma_{z_2x_1}, \sigma_{z_2x_2})$. In fact, λ_{x_1y} is not identifiable here.

Modern Extensions of Instrumental Sets:

1. Generalized Half-Trek Criterion (gHTC, Chen 2016; Weihs et al. 2017) & Auxiliary Variable Sets (AVS, Chen et al. 2017)

   *Use previously identified edges to help in identifying new ones. Existing algorithms either assumed that all parents of a node were part of the instrumental set, or enumerated all combinations of parents, making them run in **exponential** time*

2. Generalized Instrumental Sets (gIS, Brito & Pearl 2002) & Generalized AV Sets (gAVS, Chen et al. 2017):

   *Extension of Conditional IV/AV to Sets, allowing to block paths by conditioning. These methods currently have an **unknown complexity**, with existing algorithms being **exponential**.*

Identification Power & Efficiency

Algorithm	Power	Efficient?
IV	low	✓
cIV	medium	✓
IS	medium	✓
scIS	high	? → ✗
gIS	high	?
HTC	high	✓
gHTC	very high	? → ✓
cAV & AVS	very high	? → ✓
Our Method	**very high**	✓
gAVS	very high	?
TSID & gHTC	very high	? → ✗
Gröbner	complete	✗

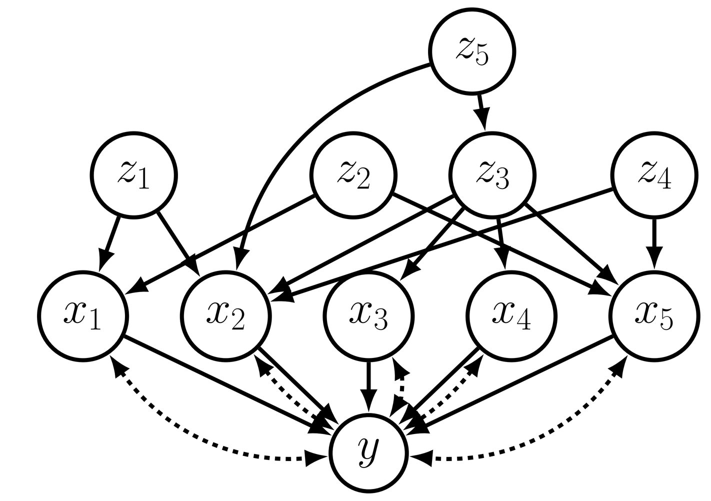
Existing methods ordered *roughly* by identification power.

→ represents methods for which we determined complexity in this work.

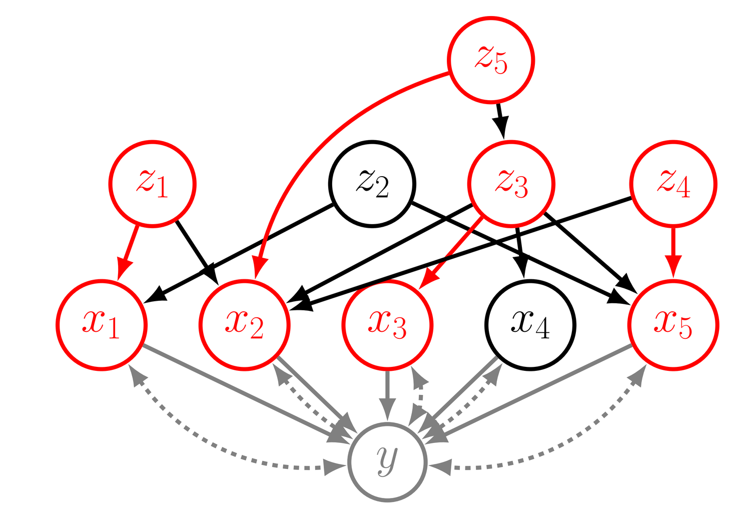
Our Approach

The Max-Match-Block Algorithm

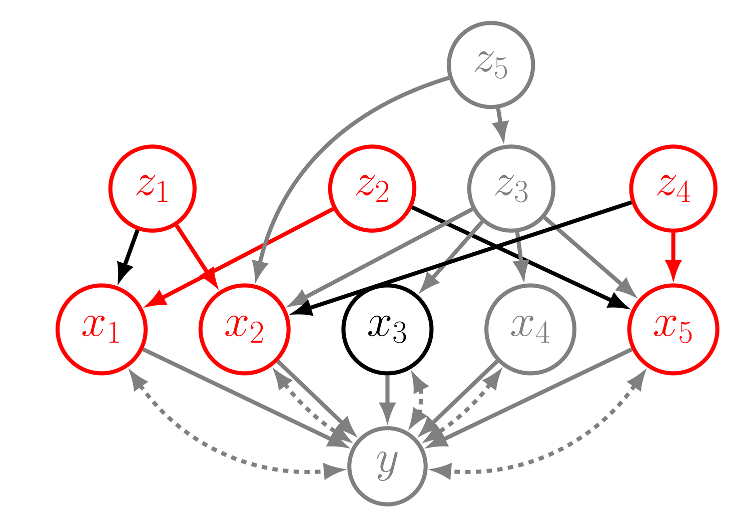
How to find the *largest* instrumental set in this graph?



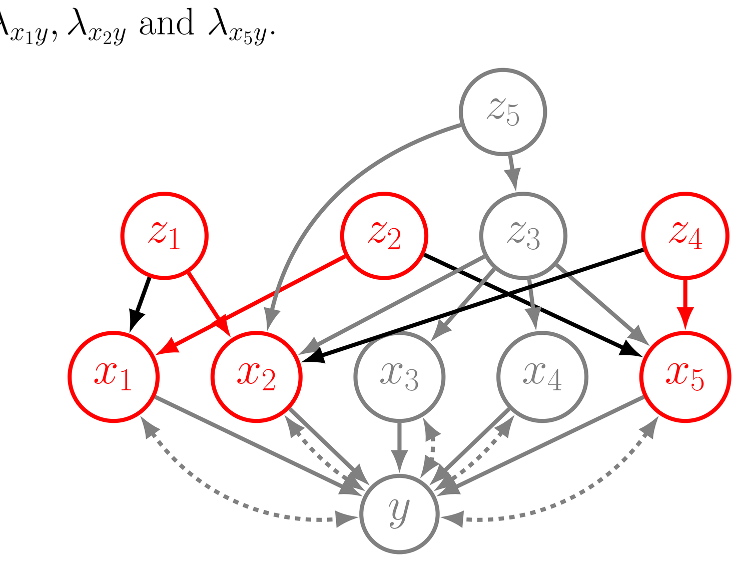
1. Run a vertex max-flow from candidate instruments (Z) to the parents of Y (X). We proved that any node in X that has no flow through it cannot be part of *any* instrumental set.



2. X_4 has no flow through it in the above graph. Disable it, and all of its ancestors, since they cannot be part of an IS either. Then, run another max-flow on the resulting graph:



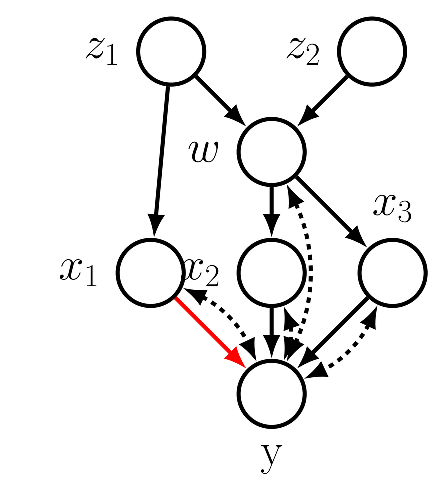
3. Now X_3 has no flow through it. Repeat the procedure one more time, leaving us with the instrumental set $\{z_1, z_2, z_4\}$ that can be used to solve for $\lambda_{x_1y}, \lambda_{x_2y}$ and λ_{x_5y}.



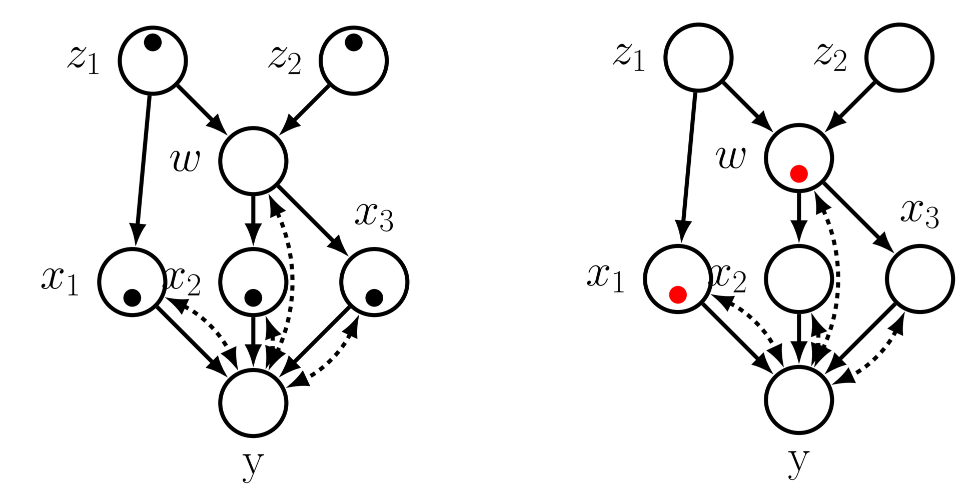
Please note that in reality, this algorithm would run on the *flow graph*, where treks are encoded explicitly. While the mechanics of the algorithm are the same, using the flow graph allows us to encode knowledge of previously solved edges (check out the *auxiliary flow graph* in our paper!).

The Instrumental Cutsets Method

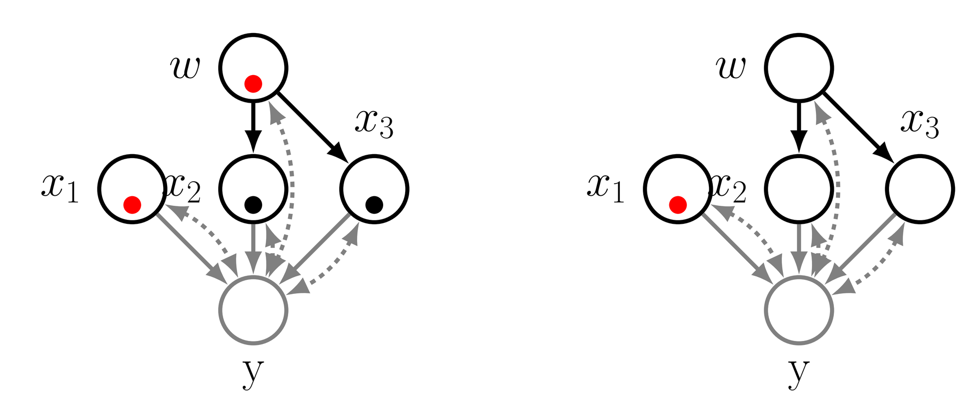
How to solve for λ_{x_1y}? This graph does not have an instrumental set, nor a conditional instrumental variable.



1. Find the vertex min-cut between Z and X closest to X (in the *flow* graph - the original graph is used for demonstration only).

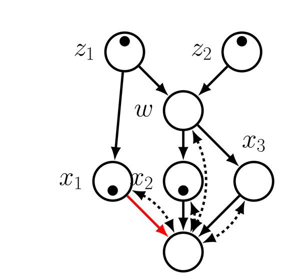


2. Run a Max-Match-Block between this min-cut and X. Only the element x_1' is match-blocked to x_1'



3. When such a match-block exists, we showed that λ_{x_1y} is identified with the equation:

$$\lambda_{x_1y} = \frac{\det\begin{pmatrix}\sigma_{z_1y} & \sigma_{z_1x_2} \\ \sigma_{z_2y} & \sigma_{z_2x_2}\end{pmatrix}}{\det\begin{pmatrix}\sigma_{z_1x_1} & \sigma_{z_1x_2} \\ \sigma_{z_2x_1} & \sigma_{z_2x_2}\end{pmatrix}}$$



Summary

We introduced a new approach for identification in linear SCMs.

1. We developed a polynomial-time algorithm to compute a max-match-block, and showed how it can be applied to make efficient versions of gHTC and AVS algorithms;
2. We proposed the Instrumental Cutset method, which is polynomial-time and subsumes state-of-the-art algorithms;
3. We showed that despite its promise, the tsIV criterion cannot be applied in polynomial-time without further assumptions.

Purdue University · Brex · Columbia University