Causal Discovery in Linear Structural Causal Models with Deterministic Relations

Yuqin Yang[1] Mohamed Nafea[2] AmirEmad Ghassami[3] Negar Kiyavash[4]

[1]Georgia Institute of Technology [2]University of Detroit Mercy [3]Johns Hopkins University [4]EPFL

Classes of linear SCMs

Linear G-SCM:

$$x = f_x(\overbrace{Pa(x)}^{\text{Parents}}) + g_x(\overbrace{S(x)}^{\text{Sources}})$$

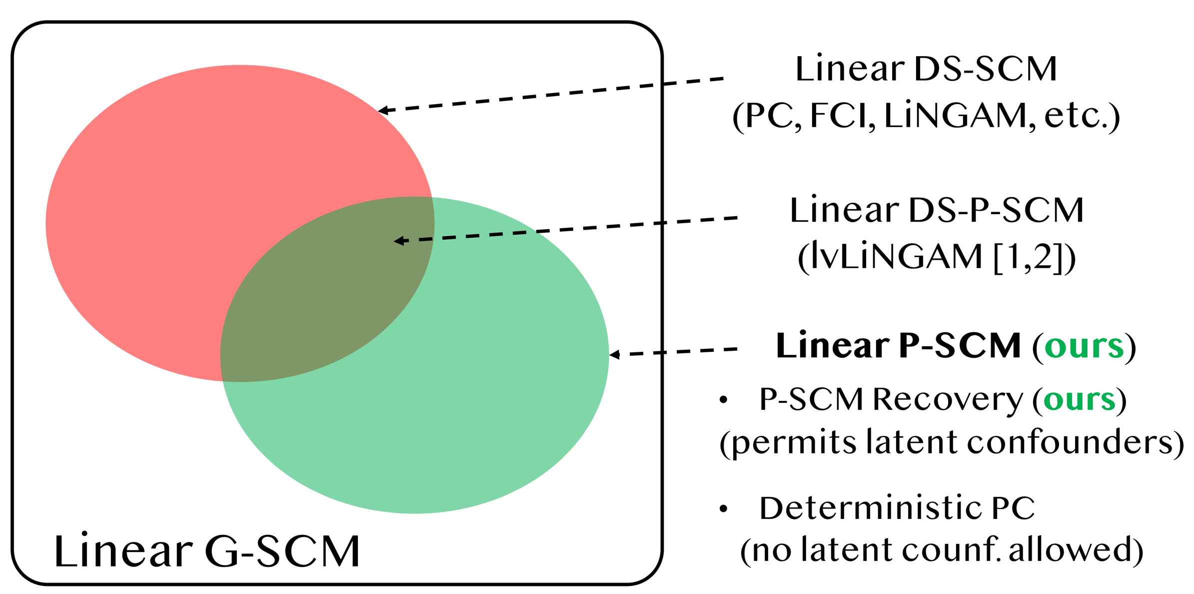
Observed variable — Linear — Not necessarily linear

Linear DS-SCM:

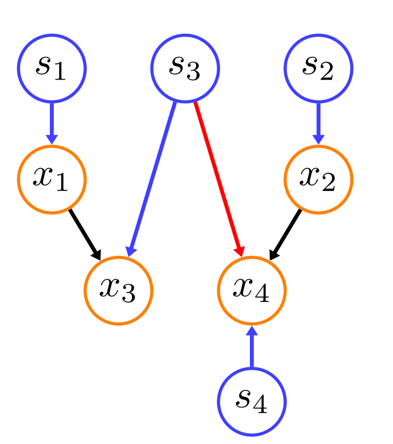
- $S(x)$ contains at least one distinct source (DS)
- No obs. variable can be a deterministic function of its obs. parents and/or latent confounders

Linear P-SCM (ours):

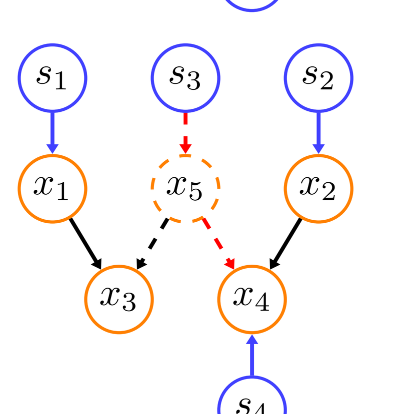
- $S(x)$ does not necess. have a distinct source
  No distinct source $\equiv$ deterministic relation
- Linear latent confounding: $g_x(S(x)) = \sum_{j=1}^{m} b_j s_j$
- Can be used to model influence propagation



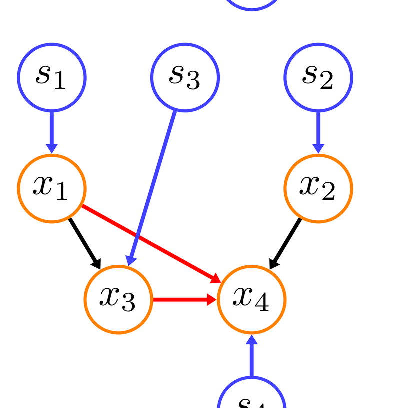
Example: Failure mode of DS-SCM



Ground-truth with 4 observed variables and 4 sources. Can be modeled as a linear P-SCM as it allows for shared exogenous connections.



Cannot be modeled as a DS-SCM with latent confounder: Assuming a latent variable x_5, x_3 has no distinct source component (i.e., deterministic).



Incorrect recovery if a DS-SCM (without latent confounders) is assumed to be the gener. model: Associating s_3 to x_3 leads to two erroneous edges.

Comparison with linear DS-SCM

P-SCM Model Assumption (weaker than DS)

Assumption 0: Every obs. variable has strictly more source components than its parents

Assumptions	DS-SCM	P-SCM
(i) Distinct source	Yes	No
(ii) Assumption 0	Yes	Yes
(iii) Linear latent confounding	No	Yes

- If Assumptions (i) and (iii) are both satisfied, then the resulting model (DS-P-SCM) is the intersection of DS-SCM and P-SCM
- Majority of works on linear causal models in fact considers DS-P-SCM; our results strictly expands the considered model space

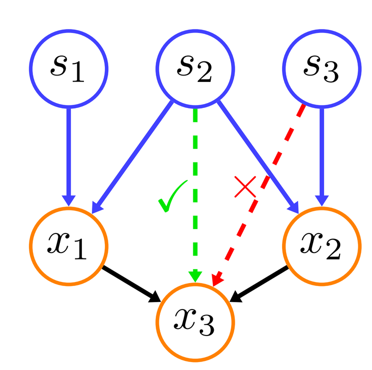
Conditions for unique identifiability

Necessary & sufficient conditions under faithfulness & source separability

Condition 1: Unique component condition

- Prevents certain exogenous connections:
  a) Unique component(s) in possible parents
  b) Shared component(s) in possible parents with no unique components

Example

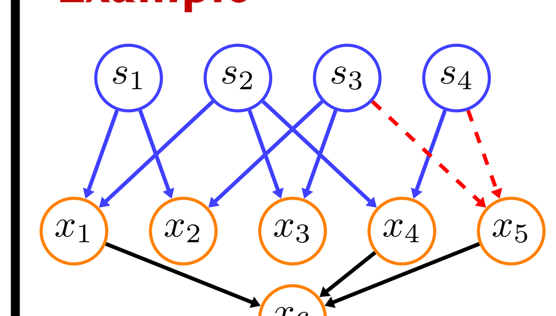


- s_3 is a u.c. of x_2; it is only connected to one parent of x_3
- If the red arrow is present, then there are 2 paths from s_3 to x_3 and we **cannot** recover either one
- If the green arrow is present, then we **can** recover a_{31}, a_{32} (using the corresponding u.c.)

Condition 2: Marriage condition

- Ensures that the possible parent set has a number of exogenous connections sufficient for recovery

Example



$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & 0 & 0 \\ b_{21} & 0 & b_{23} & 0 \\ 0 & b_{32} & b_{33} & 0 \\ 0 & b_{42} & 0 & b_{44} \\ 0 & 0 & b_{53} & b_{54} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{bmatrix}$$

- $x_5 =$ linear combination of $[x_1 : x_4]$
- The causal effect of x_5 on x_6 can be replaced by this linear combination
- This will not happen if and only if
  $\forall X \subseteq [x_1 : x_5] : |X| \leq |\text{Source components of } X|$

Main result

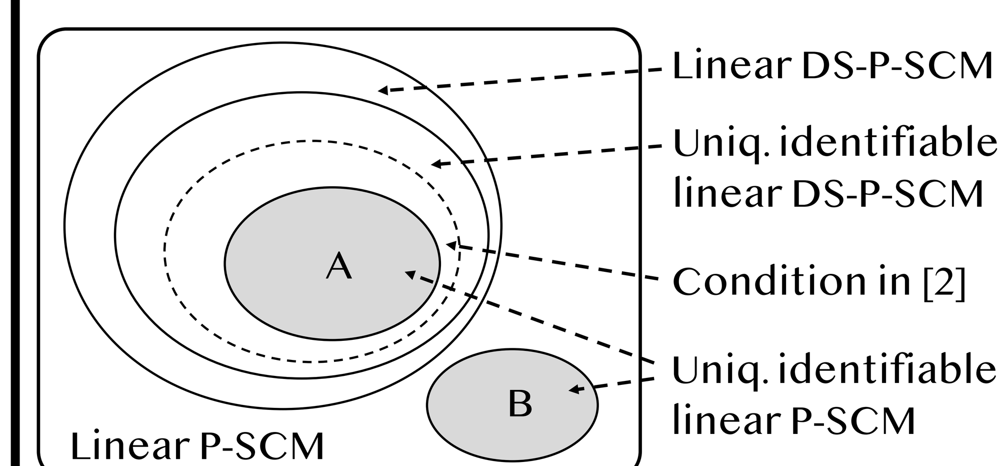
Theorem 1: Under faithfulness & source separability, a linear P-SCM is uniquely identifiable if and only if Conditions 1 and 2 hold for every observed variable.*

*Definition of unique components for general structures, and full statements of Conditions 1 & 2 can be found in Section 3 of our paper.

Reduction of the conditions

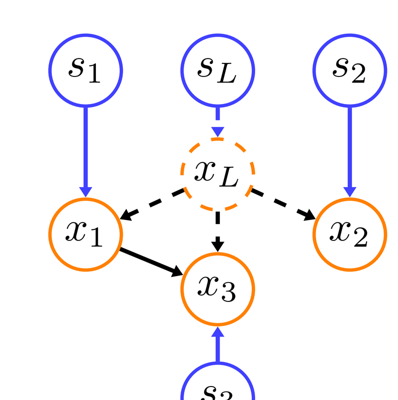
Theorem 2: Under distinct source assumption (DS-P-SCM), Conditions 1 and 2 can be reduced to Condition 3.

Condition 3: For each exogenous connection, there are either no other paths, or at least two other distinct paths from the source to the obs. variable.

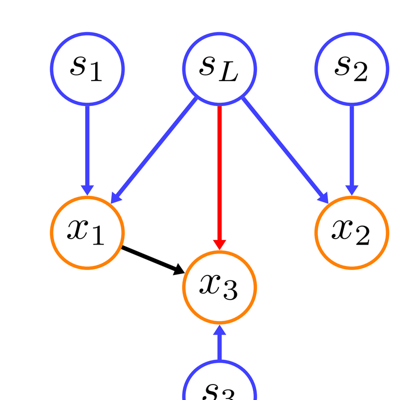


Thm. 1: $\mathcal{M} \in \mathbf{A} \cup \mathbf{B}$ ➡ Thm. 2: $\mathcal{M} \in \mathbf{A}$ $\subseteq$ Thm. 17 in [2] for DS-P-SCM

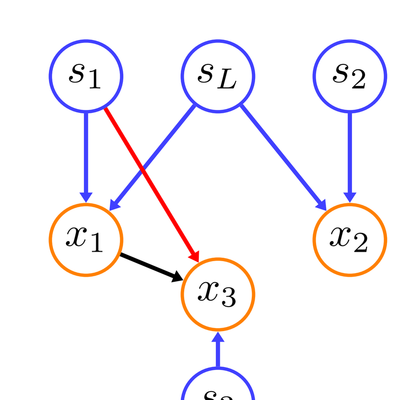
Reduced to DS-P-SCM — The search space in [2] is strictly smaller than in Thm. 2



True DS-P-SCM
- Satisfies the condition in [2]
- Uniquely identifiable



Corresponding P-SCM
- Violates Thm. 2
- Not uniq. Identifiable as a P-SCM



Observationally equivalent P-SCM
- Not a DS-P-SCM
- Does not contradict uniq. identifiability when the starting assumption is a DS-P-SCM

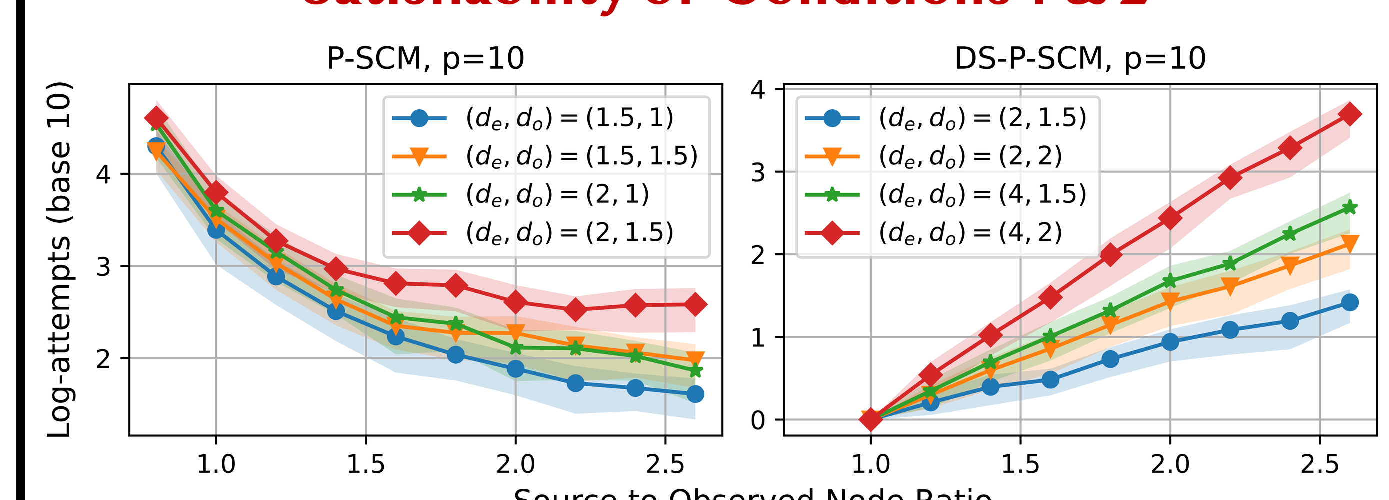
P-SCM Recovery algorithm

Input: Recovered mixing matrix $\tilde{\mathbf{W}}$.
Output: Adjacency matrix $\mathbf{A}$, exog. connection matrix $\mathbf{B}$.

1 **Initialize:** $\tilde{\mathbf{A}} = \mathbf{I}$, $\mathbf{B} = \mathbf{0}$; `// Total causal effect`
2 Repermute $\tilde{\mathbf{W}}$ s.t. the number of non-zero entries in each row is in an increasing order;
3 **for** $k = 1 : p$ **do**
4 Find possible parent set $\mathcal{P}_k$ using $\tilde{\mathbf{W}}$;
5 Compute total causal effect from $x_i \in \mathcal{P}_k$ to x_k using unique components of x_i in an iterative manner, until no more unique components can be found in the last subset $\mathcal{I}_k$;
6 Compute total causal effect from $x_i \in \mathcal{I}_k$ to x_k by solving overdetermined linear system;
7 Add remaining source components in x_k to the exogenous connections;
8 $\mathbf{A} = \mathbf{I} - \tilde{\mathbf{A}}^{-1}$;
9 Repermute matrices $\mathbf{A}$, $\mathbf{B}$ according to the reversed order from Step 2;

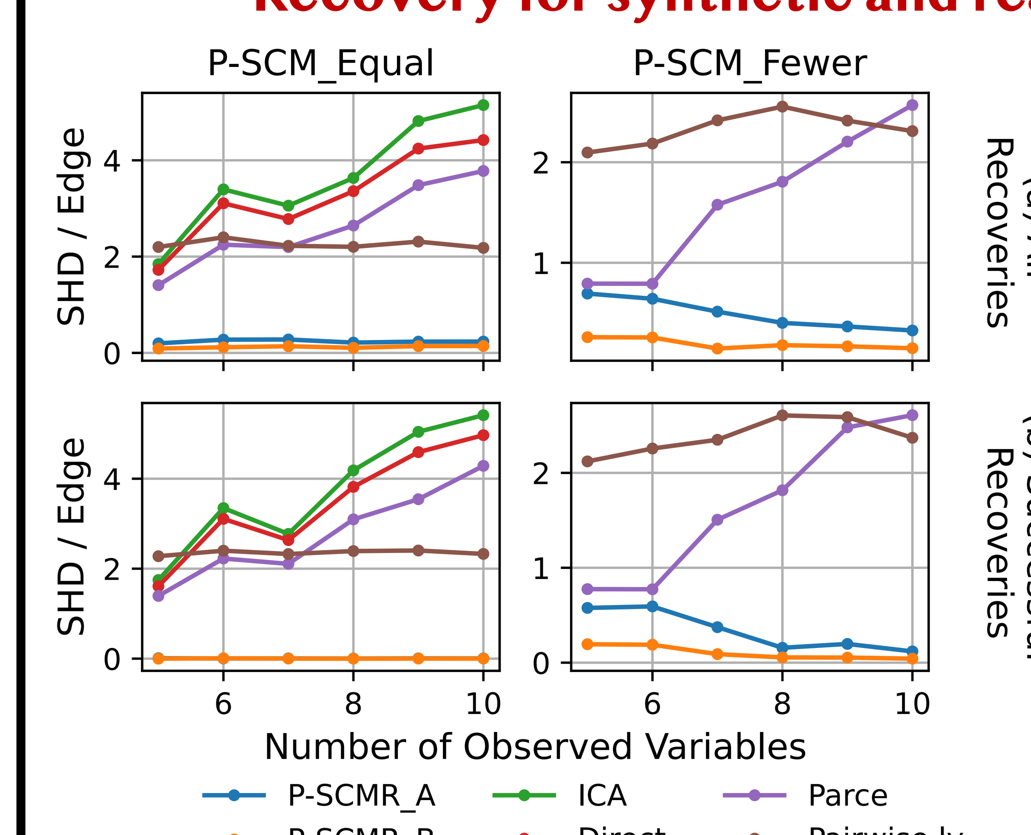
Experiments

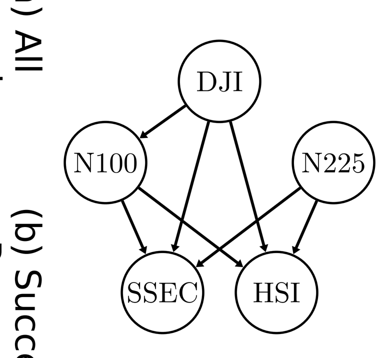
Satisfiability of Conditions 1 & 2



- Reports # attempts to randomly generate one satisfying model
- d_e/d_o: Average causal/exog. connections for each node/source

Recovery for synthetic and real data





Model the relative daily change of 5 stock indices as a linear P-SCM

- P-SCMR_A/B: Recovery of A/B using Alg. Above
- Successful recovery: Support of W is recovered

[1] Hoyer et al. "Estimation of causal effects using linear non-Gaussian causal models with hidden variables". *International Journal of Approx. Reas., 2008.*
[2] Salehkaleybar et al. "Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables". *JMLR 2020.*