Causal Discovery in Additive Noise Models using Beam Search

Hans Jarett J. Ong[1] Brian Godwin S. Lim[1,2] Renzo Roel P. Tan[1,3] Kazushi Ikeda[1]

[1]Nara Institute of Science and Technology, Nara, Japan
[2]Kyoto University, Kyoto, Japan
[3]Ateneo de Manila University, Metro Manila, Philippines

ISAROB 2026

Introduction

Motivation

- Autonomous agents (e.g., in robotics) must understand the underlying mechanics of their environment to predict the consequences of actions.
- This requires moving from correlation/patterns to **causality**.
- Direct experimentation (interventions or Randomized Controlled Trials) is straightforward but often infeasible or unethical.
- **Causal Discovery** allows us to infer cause and effect from widely available *observational data*.

Graphical Causal Models

Structural Causal Models (SCMs)

- We model the system using a Directed Acyclic Graph (DAG) $\mathcal{G}$ over variables $X = \{X_1, \ldots, X_d\}$.
- Each variable X_i is a function of its parents PA_i and an exogenous noise term N_i:

$$X_i := f_i(\text{PA}_i, N_i)$$

The Identifiability Challenge

- Without functional assumptions, we can only identify the **Markov Equivalence Class (MEC)**—a set of graphs that are statistically indistinguishable.
- *Example:* $X \to Y$ and $Y \to X$ can generate the same observational distribution in the general case.

Background: Additive Noise Models (ANMs)

To overcome the ambiguity of MECs, we restrict the functional form to the **Additive Noise Model (ANM)**.

Model Assumption

We assume the noise is **additive**:

$$X_i := f_i(\mathrm{PA}_i) + N_i$$

Crucially, the noise N_i must be statistically **independent** of the parents PA_i ($N_i \perp\!\!\!\perp \mathrm{PA}_i$).

Why this works?

- This restriction breaks the symmetry between cause and effect.
- It allows us to identify the **unique true DAG** from observational data, provided f_i is non-linear.

The Standard Approach: RESIT

The **RE**gression with **S**ubsequent **I**ndependence **T**est (RESIT) algorithm determines the causal order by iteratively finding a "Sink" node (a variable with no effects).

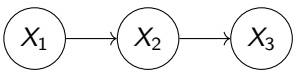
The Greedy Procedure

At each step k, given a set of remaining variables S:

1. **Candidate Evaluation:** For every variable $X_j \in S$:
   - Regress X_j on all other variables ($S \setminus \{X_j\}$).
   - Compute residuals $\hat{N}_j$.
   - Measure dependence between $\hat{N}_j$ and regressors (using HSIC).
2. **Selection (Greedy Step):** Select the variable X_{j^*} with the **minimum dependence** (smallest mutual information).
3. **Removal:** Set X_{j^*} as the sink and **remove it** from S.
4. Repeat until all variables are ordered.

RESIT: A Concrete Example

Consider a true chain: $X_1 \to X_2 \to X_3$.



Goal: Identify the Sink (X_3).

Iteration 1: Remaining $\{X_1, X_2, X_3\}$

- **Try** X_1 **as Sink:** Regress X_1 on $\{X_2, X_3\}$.
  $\to$ Residuals are **Dependent** (Effect predicts Cause).
- **Try** X_2 **as Sink:** Regress X_2 on $\{X_1, X_3\}$.
  $\to$ Residuals are **Dependent**.
- **Try** X_3 **as Sink:** Regress X_3 on $\{X_1, X_2\}$.
  $\to$ Residuals are **Independent**.

Result: Select X_3 as Sink. Remove X_3.

Remaining: $\{X_1, X_2\}$

Iteration 2:

- Regress X_2 on $X_1 \to$ **Independent** (Sink).
- Regress X_1 on $X_2 \to$ **Dependent**.

Result: Select X_2. Final Order: X_1, X_2, X_3.

The Problem: Why is Greedy Search Insufficient?

The Irreversibility Problem

- RESIT commits to the locally optimal choice at each step.
- Once a variable is removed, **it cannot be revisited**.

Sources of Error

1. **Finite Sample Noise:** With small n, a non-sink variable may appear statistically independent simply by chance.
2. **Unmeasured Confounders:** Violations of causal sufficiency create spurious associations that mislead the local independence test.

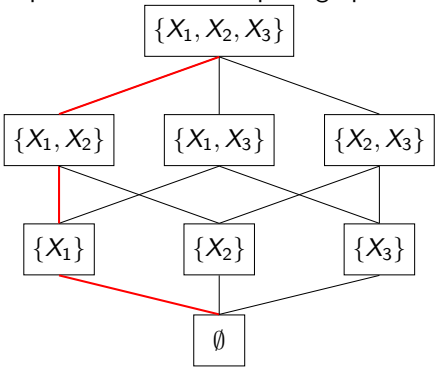
One early error propagates through the entire chain, ruining the final graph.

Proposed Method: Beam Search Optimization

We generalize RESIT by framing causal ordering as a path-search problem on a state-space graph.

Algorithm:

- **Nodes:** Sets of unordered variables.
- **Edges:** Selection of a sink variable.
- **Cost:** HSIC test statistic (Dependence).
- **Beam Width (w):** Instead of keeping 1 path (Greedy), we keep the top-w lowest cost paths.



State-space graph. Adapted from Suzuki et al. (2024).

Benefit: This approach recovers from noisy local optima by exploring multiple alternative orderings simultaneously.

Experimental Setup

Data Generation

- Non-linear ANMs generated via Gaussian Processes (Matérn Kernel).
- **Structure:** Random DAGs with average degree 1.5.

Parameters

- **Nodes (d):** 10, 30.
- **Sample Size (n):** 100, 250, 500, 1000.
- **Confounding:** 0% (Ideal), 10% (Unmeasured Confounders).
- **Beam Width (w):** 1 (Greedy) to 32.

Evaluation Metrics

1. Structural Hamming Distance (SHD)

- A purely structural assessment. Counts edge disagreements (additions, deletions, reversals) required to transform the estimated graph $\mathcal{H}$ into the true graph $\mathcal{G}$.

$$\text{SHD}(\mathcal{G}, \mathcal{H}) = |\{(i,j) : \text{edge type between } i,j \text{ differs in } \mathcal{G} \text{ and } \mathcal{H}\}| \tag{1}$$

2. Structural Intervention Distance (SID)

- Assesses the ability to correctly predict interventions. Counts pairs (X_i, X_j) where the parent set in $\mathcal{H}$ is **not** a valid adjustment set for the causal effect of X_i on X_j.

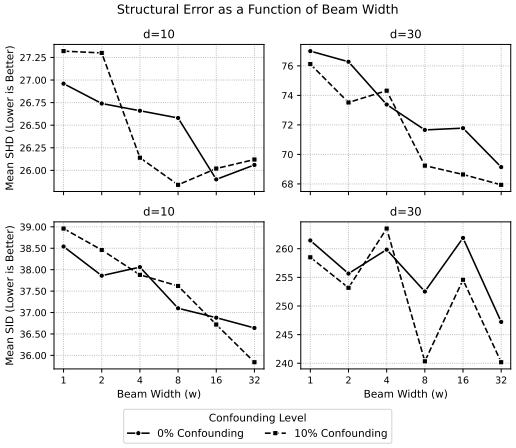
$$\text{SID}(\mathcal{G}, \mathcal{H}) = |\{(i,j) : p_{\mathcal{H}}(X_j|do(X_i)) \neq p_{\mathcal{G}}(X_j|do(X_i))\}| \tag{2}$$

- **Lower is Better:** A lower SID implies higher reliability for causal prediction.

Results: Structural Error vs. Beam Width

Figure 1 shows the mean structural error ($n = 250$).

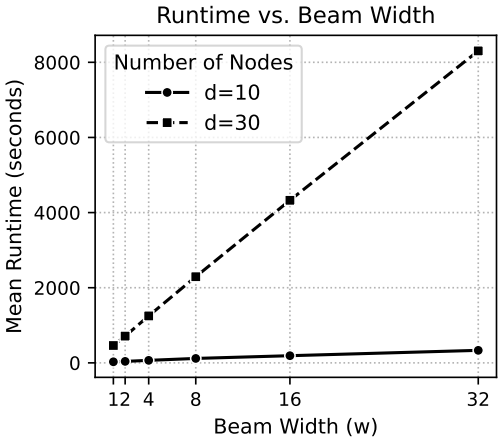
- Performance generally improves as w increases.
- Improvement is most pronounced for high-dimensional ($d = 30$) and confounded cases.



Results: Computational Cost

Figure 2 shows the runtime analysis.

- Runtime scales approximately linearly with beam width w.
- This represents a manageable trade-off: significant accuracy gains for a feasible increase in computation.



Detailed Comparison ($n = 250$)

Comparison of Greedy ($w = 1$) vs. Beam Search ($w = 32$) over 50 trials.

		SHD (↓)			**SID** (↓)		
d	Conf.	Greedy	Beam	Improv.	Greedy	Beam	Improv.
10	0%	27.0 ± 5.2	26.1 ± 5.0	3.3%	38.5 ± 6.9	36.6 ± 7.3	4.9%
	10%	27.3 ± 5.7	26.1 ± 6.0	4.4%	39.0 ± 5.7	$\mathbf{35.8 \pm 7.9}^{*}$	**8.0%**
30	0%	77.0 ± 12.4	$\mathbf{69.1 \pm 11.7}^{***}$	**10.2%**	261.4 ± 51.8	247.2 ± 45.1	5.4%
	10%	76.1 ± 13.4	$\mathbf{67.9 \pm 12.9}^{**}$	**10.7%**	258.5 ± 52.3	$\mathbf{240.2 \pm 48.8}^{*}$	**7.1%**

Table: Statistically significant improvements observed in complex ($d = 30$) and confounded scenarios.

Robustness to Sample Size

SID Improvement (%) of Beam ($w = 32$) over Greedy ($w = 1$) across varying n.

d	Conf.	**SID Improvement (%)** by Sample Size n			
		100	250	500	1000
10	0%	1.8%	4.9%	**5.2%***	−0.5%
	10%	3.0%	**8.0%***	0.0%	1.1%
30	0%	−1.4%	5.4%	**4.7%****	0.7%
	10%	1.1%	**7.1%***	**6.7%*****	**1.4%****

Key Takeaways:

- **Intermediate Regime ($n = 250, 500$):** Highest gains. The data is noisy enough to trap greedy search, but informative enough for Beam search to find the signal.
- **Large Sample ($n = 1000$):** Performance converges as local optima align with global optima.

Conclusion

Summary

- **Identified Weakness:** Greedy search (Standard RESIT) is brittle in high-variance settings (finite samples, confounders) due to irreversible steps.
- **Proposed Solution:** Generalized RESIT to a **Beam Search** framework.
- **Key Result:** Significant improvements in structural accuracy (SHD/SID) in intermediate sample regimes and high-dimensional graphs.
- **Practicality:** Offers a tunable trade-off between speed and robustness.

Thank You