Optimal estimation of linear non-Gaussian structure equation models

Sunmin Oh*, Seungsu Han, Gunwoong Park

Seoul National University
Department of Statistics

Main Objectives and Outline

Main Objectives

- Establish the optimal sample complexity for learning Linear non-Gaussian acyclic models (LiNGAM), $n = \Theta\left(d_{in} \log \frac{p}{d_{in}}\right)$.
- Develop a LiNGAM learning algorithm using distance covariance that achieves the optimal sample complexity without assuming (parental) faithfulness or a known indegree.

- Introduction
- Recent Works on LiNGAMs
- New Properties of LiNGAMs
- Proposed Algorithm
- Numerical Experiments & Real Data Analysis

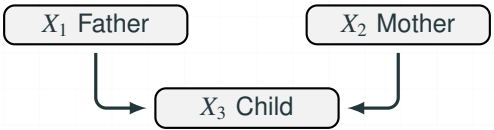
Introduction

Directed Acyclic Graphical Models

By recovering the directed acyclic graph (DAG) $\mathcal{G}^*$ from the data $\mathcal{D}$,

- we can learn the conditional dependency structure between variables.
- we can estimate a statistical model of the underlying distribution. In this case, the model can be used to generalize new instances.

Suppose that there are three variables of family gene information, $X_3 = f(X_1, X_2)$ (functional):



$$X_1 \perp\!\!\!\perp X_2, \quad X_1 \not\perp\!\!\!\perp X_2 \mid X_3, \quad X_1 \not\perp\!\!\!\perp X_3, \quad X_1 \not\perp\!\!\!\perp X_3 \mid X_2, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_2 \not\perp\!\!\!\perp X_3 \mid X_1.$$

Definitions

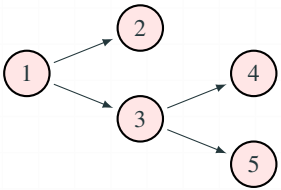


Figure 1: Directed graph G

- Parent (Pa): $1 \to \{2, 3\}$ and $\text{Pa}(3) = \{1\}$.
- Child (Ch): $3 \to \{4, 5\}$ and $\text{Ch}(3) = \{4, 5\}$.
- Descendant (De): $\text{De}(1) = \{2, 3, 4, 5\}$.
- Non-descendant (Nd): $\text{Nd}(3) = \{1, 2\}$, $\text{Nd}(5) = \{1, 2, 3, 4\}$.
- Topological layers ($\mathcal{A}_t$): containing nodes whose longest distance to a source node is t.

$$\mathcal{A}_0 = \{1\},\ \mathcal{A}_1 = \{2, 3\},\ \mathcal{A}_2 = \{4, 5\}.$$

  - For each $j \in \mathcal{A}_r$, $\text{Pa}(j) \subset \cup_{t=0}^{r-1} \mathcal{A}_t \subset \text{Nd}(j)$.
  - Ordering (π): $\pi = (1, 2, 3, 4, 5)$ or $\pi = (1, 3, 2, 5, 4)$.
- Maximum indegree (d_{in}): maximum number of incoming edges, $d_{in} = 1$.

Definition of Linear SEMs

Linear Structural Equation Models

A Linear SEM is a DAG model where each variable is expressed as a linear function of its parents variables plus an independent error.

$$X_j = \sum_{k \in \text{Pa}(j)} \beta_{k,j} X_k + \epsilon_j, \qquad \forall j = 1, \ldots, p,$$

$$(X_1, X_2, ..., X_p)^\top = B(X_1, X_2, ..., X_p)^\top + (\epsilon_1, \epsilon_2, ..., \epsilon_p)^\top.$$

where $[B]_{j,k} \neq 0$ for all $k \in \text{Pa}(j)$, otherwise $[B]_{j,k} = 0$.

- The error distributions in sub-Gaussian LiNGAM models are continuous and sub-Gaussian, but explicitly non-Gaussian.
  - If error distributions are uniform, then it is a sub-Gaussian LiNGAM.

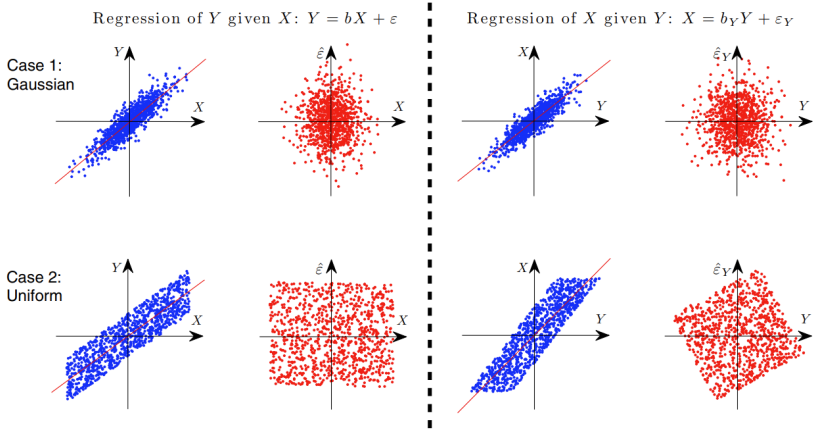
Recent Works on LiNGAMs

Model Identifiability

A critical issue of graphical models: *Identifiability*

- We can distinguish G_2 and G_3 from G_1.
- We cannot identify the direction of an edge. Hence, we cannot distinguish G_2 from G_3.
- Existing algorithms recover the skeleton or CPDAG. (e.g., PC, GES algorithms)

Model Identifiability of LiNGAMs



- The direction of an edge between two variables can be determined by assessing the dependency between (residualized) variables.

Ordering Recovery of LiNGAMs

$$X_1 = \epsilon_1, \quad X_2 = \beta_1 X_1 + \epsilon_2 = \beta_1 \epsilon_1 + \epsilon_2, \quad X_3 = \beta_2 X_2 + \epsilon_3 = \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3,$$

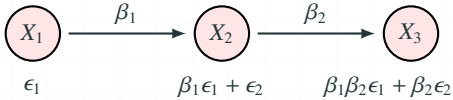
where ϵ_j's are independent and non-Gaussian.

(1) Choose an exogenous variable based on non-Gaussianity and independence. $\rightarrow X_1 = \epsilon_1$.

(2) Remove the effect of the exogenous variable from the other variables.
$\rightarrow e_{2,\{1\}} = X_2 - \Sigma_{2,1}(\Sigma_{1,1})^{-1}X_1 = \epsilon_2$, $e_{3,\{1\}} = X_3 - \Sigma_{3,1}(\Sigma_{1,1})^{-1}X_1 = \beta_2\epsilon_2 + \epsilon_3$.

(3) The iteration of effect removal and identification of exogenous variables recovers the ordering of a graph.
$\rightarrow e_{2,\{1\}} = \epsilon_2$ becomes an exogenous variable, whereas $e_{3,\{1\}} = \beta_2\epsilon_2 + \epsilon_3$ does not.

Details of Ordering Recovery of LiNGAMs



$X_1 = \epsilon_1$	$e_{2,\{1\}} = \epsilon_2$	$e_{3,\{1\}} = \beta_2\epsilon_2 + \epsilon_3$
$e_{1,\{2\}} = (1 - \Sigma_{1,2}(\Sigma_{2,2})^{-1}\beta_1)\epsilon_1 - \Sigma_{1,2}(\Sigma_{2,2})^{-1}\epsilon_2$	$X_2 = \beta_1\epsilon_1 + \epsilon_2$	$e_{3,\{2\}}$
$e_{1,\{3\}}$	$e_{2,\{3\}}$	$X_3 = \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3$

↓

$e_{2,\{1\}} = \epsilon_2$	$e_{3,\{1\}} - \frac{\mathrm{Cov}(e_{2,\{1\}}, e_{3,\{1\}})}{\mathrm{Var}(e_{2,\{1\}})} e_{2,\{1\}} = \beta_2\epsilon_2 + \epsilon_3 - \beta_2\epsilon_2 = \epsilon_3$
$e_{2,\{1\}} - \frac{\mathrm{Cov}(e_{3,\{1\}}, e_{2,\{1\}})}{\mathrm{Var}(e_{3,\{1\}})} e_{3,\{1\}} = \epsilon_2 - \frac{\mathrm{Cov}(e_{3,\{1\}}, e_{2,\{1\}})}{\mathrm{Var}(e_{3,\{1\}})}(\beta_2\epsilon_2 + \epsilon_3)$	$e_{3,\{1\}} = \beta_2\epsilon_2 + \epsilon_3$

✓ For the algorithm, a least squares regression with $p - 1$ predictors is needed, making $n < p$ infeasible.

Strategies for the scenario when $n < p$ (1)

$$X_1 = \epsilon_1, \quad X_2 = \beta_1 X_1 + \epsilon_2 = \beta_1 \epsilon_1 + \epsilon_2, \quad X_3 = \beta_2 X_2 + \epsilon_3 = \beta_1 \beta_2 \epsilon_1 + \beta_2 \epsilon_2 + \epsilon_3,$$

where ϵ_j's are independent and non-Gaussian.

Wang and Drton (2020)

- This study interprets the effect removal step as the elimination of confounding effects.
- By identifying a set of variables that eliminates all confoundings between variables, the ordering is recovered in a top-down manner.
- The assumption of parental faithfulness is required.

Strategies for the scenario when $n < p$ (2)

$$X_1 = \epsilon_1, \quad X_2 = \beta_1 X_1 + \epsilon_2 = \beta_1 \epsilon_1 + \epsilon_2, \quad X_3 = \beta_2 X_2 + \epsilon_3 = \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3,$$

where ϵ_j's are independent and non-Gaussian.

Zhao et al. (2022)

- The ordering is recovered in a bottom-up manner by utilizing the independence of the exogenous term of terminal nodes from all preceding variables.
  - ▹ $\epsilon_3 \perp\!\!\!\perp X_1$ and $\epsilon_3 \perp\!\!\!\perp X_2$, but $\epsilon_2 \not\perp\!\!\!\perp X_3$.
- The exogenous term is revealed by projecting each variable onto its Markov blanket.

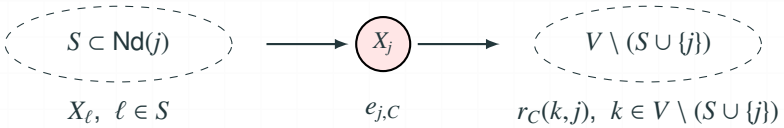
New Properties of LiNGAMs

Expanding Scope of Independence Analysis

For $j, k \in V$ and $C \subset V \setminus \{j, k\}$, denote the residuals as

- $e_{j,C} = X_j - \Sigma_{j,C}(\Sigma_{C,C})^{-1} X_C$.
- $r_C(j,k) = e_{j,C} - \frac{\mathrm{Cov}(e_{j,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})} e_{k,C}$.

Examine the independence relationships among (residualized) variables for $C \subset S$.



Optimizing Exogeneity Identification

Proposition 1

Let $P(X)$ be generated from a LiNGAM with DAG G and the topological layers $\cup_{t=0}^{T-1} \mathcal{A}_t$. For any $r \in \{1, 2, ..., T-1\}$, $j \in \mathcal{A}_r$, $k \in V \setminus \cup_{t=0}^{r} \mathcal{A}_t$, and $S_{r-1} = \cup_{t=0}^{r-1} \mathcal{A}_t$,

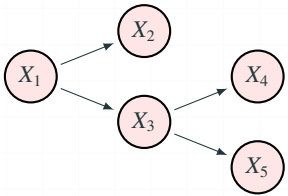
(i) there exists $C \subset S_{r-1}$ satisfying $\text{Pa}(j) \subset C$, and then $e_{j,C} \perp\!\!\!\perp r_C(\ell, j)$ for all $\ell \in V \setminus (S_{r-1} \cup \{j\})$.

(ii) if $C \subset S_{r-1}$ satisfies $e_{j,C} \perp\!\!\!\perp X_\ell$ for all $\ell \in S_{r-1}$, then $\text{Pa}(j) \subset C$.

(iii) for any $C \subset S_{r-1}$, there exists $\ell \in V \setminus (S_{r-1} \cup \{k\})$ such that $e_{k,C} \not\perp\!\!\!\perp r_C(\ell, k)$.

- Topological layers of a LiNGAM can be recovered iteratively by (i) and (iii) in a top-down manner.
  - The set of variables required to make a variable exogenous is reduced to its parent set rather than the entire set of preceding elements in the causal ordering or Markov blanket.
- By (ii), the minimal conditioning set $C \subset S_{r-1}$, for which $e_{j,C}$ is independent of X_ℓ for all $\ell \in S_{r-1}$, is $\text{Pa}(j)$.
  - Notably, (ii) does not rely on the assumption of parental faithfulness.

Example



ϵ_j are independent and non-Gaussian.

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \beta_{1,2} X_1 + \epsilon_2 \\
X_3 &= \beta_{1,3} X_1 + \epsilon_3 \\
X_4 &= \beta_{3,4} X_3 + \epsilon_4 \\
X_5 &= \beta_{3,5} X_3 + \epsilon_5
\end{aligned}
$$

- Since X_1 is the only exogenous variable, $\mathcal{A}_0 = \{1\}$.
- For $r = 1$,

  (i) $\mathrm{Pa}(2) = \mathrm{Pa}(3) = S_0 = \{1\}$.

  $\Rightarrow e_{2,\{1\}} = \epsilon_2$ and $e_{3,\{1\}} = \epsilon_3$ are exogenous.

  $\Rightarrow e_{2,\{1\}} \perp\!\!\!\perp r_{\{1\}}(\ell, 2)$ for $\ell \in \{3, 4, 5\}$, and similarly for $e_{3,\{1\}}$.

  (iii) $\mathrm{Pa}(4) = \mathrm{Pa}(5) = \{3\} \not\subset S_0$.

  $\Rightarrow e_{4,\{1\}} = \beta_{3,4}\epsilon_3 + \epsilon_4$ is dependent of

  $$r_{\{1\}}(3,4) = \epsilon_3 - \frac{\beta_{3,4}\mathrm{Var}(\epsilon_3)}{\beta_{3,4}^2 \mathrm{Var}(\epsilon_3) + \mathrm{Var}(\epsilon_4)}(\beta_{3,4}\epsilon_3 + \epsilon_4),$$

  and similarly for $e_{5,\{1\}}$.

  (ii) $e_{2,\{1\}} \perp\!\!\!\perp X_1$, $e_{3,\{1\}} \perp\!\!\!\perp X_1$ and $\mathrm{Pa}(2) = \mathrm{Pa}(3) = \{1\}$.

Topological Layer Recovery

Theorem 2

Let $P(X)$ be generated from a LiNGAM with DAG G and the topological layers $\cup_{t=0}^{T-1}\mathcal{A}_t = V$. Consider any $r \in \{1, 2, ..., T-1\}$ and $S_{r-1} = \cup_{t=0}^{r-1}\mathcal{A}_t$. Then

$$\mathcal{A}_0 = \left\{j \in V : X_j \perp\!\!\!\perp e_{k,\{j\}} \text{ for all } k \in V \setminus \{j\}\right\}, \text{ and}$$

$$\mathcal{A}_r = \{j \in V \setminus S_{r-1} : \exists C_j \subset S_{r-1} \text{ s.t. } X_k \perp\!\!\!\perp e_{j,C_j} \text{ for all } k \in S_{r-1} \text{ and } e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\}.$$

Moreover, for each $j \in \mathcal{A}_r$ and the corresponding set

$$C_j = \{C \subset S_{r-1} : X_k \perp\!\!\!\perp e_{j,C} \text{ for all } k \in S_{r-1} \text{ and } e_{j,C} \perp\!\!\!\perp r_C(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\},$$

$\text{Pa}(j) \subset C \subset \text{Nd}(j)$ for any $C \in C_j$.

- Theorem 2 also ensures that the true ordering can be correctly identified, by substituting $\pi_r \in \pi$ as a true ordering and setting $S_{r-1} = \{\pi_1, ..., \pi_{r-1}\}$ with $S_0 = \emptyset$.

Proposed Algorithm

Dependency Score

For $j \in V \setminus \mathcal{R}$, define a dependency score relative to the preceding elements in the ordering $\mathcal{R}$.

$$\widehat{\mathcal{S}}(j, C) := \#\{k \in \mathcal{R} : \hat{e}_{j,C} \not\perp\!\!\!\perp_{\text{test}} \mathbf{x}_k\} + \#\{\ell \in V \setminus (\mathcal{R} \cup \{j\}) : \hat{e}_{j,C} \not\perp\!\!\!\perp_{\text{test}} \hat{r}_C(\ell, j)\},$$

where

$$\hat{e}_{j,C} = \mathbf{x}_j - \mathbf{x}_C(\widehat{\Sigma}_{C,C})^{-1}\widehat{\Sigma}_{Cj}, \quad \hat{r}_C(j,k) = \hat{e}_{j,C} - \frac{\widehat{\Sigma}_{j,k} - \widehat{\Sigma}_{j,C}(\widehat{\Sigma}_{C,C})^{-1}\widehat{\Sigma}_{C,k}}{\widehat{\Sigma}_{k,k} - \widehat{\Sigma}_{k,C}(\widehat{\Sigma}_{C,C})^{-1}\widehat{\Sigma}_{C,k}}\hat{e}_{k,C}.$$

- $\not\perp\!\!\!\perp_{\text{test}}$ indicates that independence is tested. Any valid independence test can be applied.
  - (e.g.) Hilbert-Schmidt independence criterion, distance covariance measure.
- $\exists C \subset \mathcal{R}$ for which $\widehat{\mathcal{S}}(j, C) = 0$.

  $\Leftrightarrow$ j is a source node in the subgraph obtained after removing $\mathcal{R}$ by Proposition 1 (i) and (iii).

  $\Rightarrow$ $\text{Pa}(j) \subset C \subset \mathcal{R}$ by Proposition 1 (ii).

OptLiNGAM Algorithm

OptLiNGAM Algorithm

Input: n i.i.d. samples $\mathbf{x}^{1:n}$ and significance level α.

Output: Estimated graph, $\widehat{G} = (V, \widehat{E})$.

Step (1): Source nodes estimation.

$$\widehat{\mathcal{A}}_0 = \left\{ j \in V : \mathbf{x}_j \perp\!\!\!\perp_{\text{test}} \hat{e}_{k,\{j\}} \text{ for all } k \in V \setminus \{j\} \right\}.$$

Step (2): Directed edges (parent) estimation

Initialize $\mathcal{R} = \widehat{\mathcal{A}}_0$.

While $V \setminus \mathcal{R} \neq \emptyset$:

For $q \in \{1, 2, ..., |\mathcal{R}|\}$:

Set $\mathcal{R}_0 = \emptyset$.

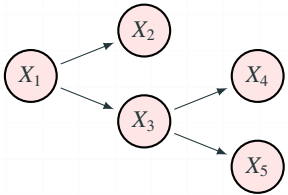
For $j \in V \setminus \mathcal{R}$:

If $\min_{|C|=q, C \subset \mathcal{R}} \widehat{S}(j, C) = 0$: Update $\mathcal{R}_0 = \mathcal{R}_0 \cup \{j\}$ and $\widehat{\text{Pa}}(j) = \arg\min_{|C|=q, C \subset \mathcal{R}} \widehat{S}(j, C)$.

If $\mathcal{R}_0 \neq \emptyset$: Update $\mathcal{R} = \mathcal{R} \cup \mathcal{R}_0$ and **Break**.

Return: $\widehat{E} = \{(k, j) : j \in V, k \in \widehat{\text{Pa}}(j)\}$.

Details of the OptLiNGAM Algorithm (1)



ϵ_j are independent and non-Gaussian.

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \beta_{1,2} X_1 + \epsilon_2 \\
X_3 &= \beta_{1,3} X_1 + \epsilon_3 \\
X_4 &= \beta_{3,4} X_3 + \epsilon_4 \\
X_5 &= \beta_{3,5} X_3 + \epsilon_5
\end{aligned}
$$

- Since X_1 is the only exogenous variable, $\mathcal{A}_0 = \{1\}$.
- $e_{2,\{1\}} = \epsilon_2$ and $e_{3,\{1\}} = \epsilon_3$ are exogenous.

$$\Rightarrow \mathcal{S}(2, \{1\}) = \mathcal{S}(3, \{1\}) = 0.$$

- $e_{4,\{1\}} = \beta_{3,4}\epsilon_3 + \epsilon_4$ and $e_{5,\{1\}} = \beta_{3,5}\epsilon_3 + \epsilon_5$, and they are dependent of

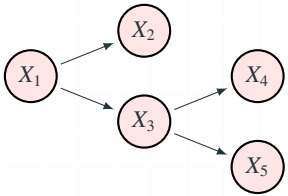
$$r_{\{1\}}(3,4) = \epsilon_3 - \frac{\beta_{3,4}\mathrm{Var}(\epsilon_3)}{\beta_{3,4}^2\mathrm{Var}(\epsilon_3) + \mathrm{Var}(\epsilon_4)}(\beta_{3,4}\epsilon_3 + \epsilon_4), \text{ and}$$

$$r_{\{1\}}(3,5) = \epsilon_3 - \frac{\beta_{3,5}\mathrm{Var}(\epsilon_3)}{\beta_{3,5}^2\mathrm{Var}(\epsilon_3) + \mathrm{Var}(\epsilon_5)}(\beta_{3,5}\epsilon_3 + \epsilon_5).$$

$$\Rightarrow \mathcal{S}(4, \{1\}), \mathcal{S}(5, \{1\}) \geq 1.$$

- $\mathcal{A}_1 = \{2, 3\}$ and $\mathrm{Pa}(2) = \mathrm{Pa}(3) = \{1\}$.

Details of the OptLiNGAM Algorithm (2)



ϵ_j are independent and non-Gaussian.

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \beta_{1,2} X_1 + \epsilon_2 \\
X_3 &= \beta_{1,3} X_1 + \epsilon_3 \\
X_4 &= \beta_{3,4} X_3 + \epsilon_4 \\
X_5 &= \beta_{3,5} X_3 + \epsilon_5
\end{aligned}
$$

- Note that

$$
\begin{aligned}
e_{4,\{3\}} = e_{4,\{2,3\}} = e_{4,\{1,3\}} = e_{4,\{1,2,3\}} = \epsilon_4, \\
e_{5,\{3\}} = e_{5,\{2,3\}} = e_{5,\{1,3\}} = e_{5,\{1,2,3\}} = \epsilon_5.
\end{aligned}
$$

- We have $\mathcal{S}(4, \{3\}) = \mathcal{S}(5, \{3\}) = 0$.
- The minimal set $\{3\}$ is identified as $\mathrm{Pa}(4) = \mathrm{Pa}(5) = \{3\}$.

✓ The number of predictors required for projection decreases from $p - 1$ or d to $d_{in} + 1$.

Theoretical Results

Theorem 5: Consistency of the OptLiNGAM Algorithm

Consider a sub-Gaussian LiNGAM with $d_{in} \leq \frac{p}{2}$. The proposed algorithm utilizes a distance covariance-based independence test a the significance level of $\alpha_n = 2(1 - \Phi(\sqrt{n\epsilon}))$, where $\epsilon \in (0, \tau_1/2)$. Then, under regularity conditions, there exist positive $A_1 > 0$ and $A_2 > 0$ such that

$$P(\widehat{G} = G) \geq 1 - A_1(p/d_{in})^{d_{in}} \exp(-A_2 n).$$

Lemma 6: Lower Bound of Sample Complexity for Arbitrary Estimator

For any $0 < \delta < 1/2$, there are positive $B_1 > 0$ and $B_2 > 0$ such that for any estimator $\widehat{G}$,

$$n \leq (1 - 2\delta) B_1 d_{in} \log(p/d_{in}) \implies \sup_{F \in \mathcal{F}_{p,d_{in}}} P(\widehat{G} \neq G(F)) \geq \delta - \frac{B_2}{p d_{in} \log(p/d_{in})},$$

where $\mathcal{F}_{p,d_{in}}$ is a class of p-node sub-Gaussian LiNGAMs with d_{in}, which satisfy regularity conditions.

Corollary 7: Optimality of the OptLiNGAM Algorithm

The proposed algorithm is optimal in sample complexity $n \asymp d_{in} \log(p/d_{in})$ under regularity conditions.

Comparison to Existing Algorithms

Algorithm	Identifiability	Sample complexity	Assumption
OptLiNGAM	Non-Gaussianity	$\Omega(d_{in}\log(p/d_{in}))$	-
TL	Non-Gaussianity	$\Omega(T^{c_1}d^6\log(p)^{c_2})$	Incoherence
MDirect	Non-Gaussianity	$\Omega((\log p)^{2K})$	Parental faithfulness
OptFGSM	Forward condition	$\Omega(d_{in}\log(p/d_{in}))$	Knowledge of d_{in}
HLSM	Backward condition	$\Omega(d^2\log p)$	Incoherence
LISTEN	Backward condition	$\Omega(d^2\log p)$	Incoherence

TL (Zhao et al., 2022), MDirect (Wang and Drton, 2020), OptFGSM (Gao et al., 2022),

HLSM (Park et al., 2021), LISTEN (Ghoshal and Honorio, 2018).

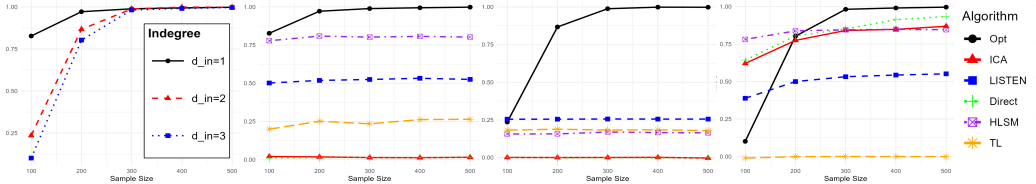
- The current sample complexity for LiNGAMs surpasses that of Gaussian SEMs, limiting its application in fields where high-dimensional causal discovery is challenged by very limited sample sizes.
- OptLiNGAM is the most sample efficient among LiNGAM learning algorithms without assuming faithfulness or a known indegree.
- This is the first result to establish optimal sample complexity for high-dimensional LiNGAMs, with an upper bound optimal up to constant factors.

Numerical Experiments

Simulation Settings

- Three types of hub graphs are generated with $d_{in} \in \{1, 2, 3\}$.
  - The number of hub nodes corresponds to d_{in}.
  - $\lfloor \log p \rfloor$ nodes are isolated, and the remaining nodes are children of the hub nodes.
- The data generation process is repeated 30 times.
  - Error terms are drawn from Beta$(0.5, 0.5)$.
  - Nonzero edge weights are sampled uniformly from $[-1.5, -0.5] \cup [0.5, 1.5]$.
- All algorithms are evaluated using the Matthews correlation coefficient (MCC), which measures the accuracy of the estimated directed edges.
  - $+1$ indicates a perfect prediction, 0 represents an average random prediction, and -1 signifies an inverse prediction.

Verification of Consistency



(a) OptLiNGAM (b) $d_{in} = 1$ (c) $d_{in} = 2$ (d) $d_{in} = 3$

Figure 1: Average MCC for 50-node sparse hub graphs by varying sample size $n \in \{100, 200, ..., 500\}$.

- As n increases, the average MCC converges to 1, signifying perfect graph recovery.
- OptLiNGAM requires fewer samples to recover sparse graphs with lower maximum indegree.
- OptLiNGAM significantly outperforms other methods as sample size increases, underscoring its advantages: being sample optimal and not requiring restrictive conditions.

Real Data Analysis

General Social Survey Data

- General Social Survey of U.S. adults(http://www.norc.org/GSS+Website) with six variables:

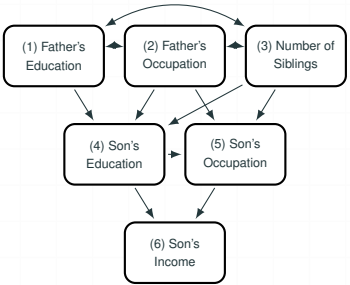
  (1) Father's Education, (2) Father's Occupation, (3) Number of Siblings,
  (4) Son's Education, (5) Son's Occupation, and (6) Son's Income.

- Duncan et al. (1972) has provided the true graph based on domain knowledge.

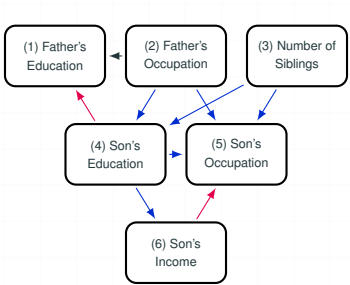
- This data was analyzed by DirectLiNGAM in Shimizu et al. (2011), specifically focused on samples for 45 years, from 1972 to 2006.

- This analysis employs data for five years, from 2002 to 2006, consisting of 355 observations, to demonstrate the ability of OptLiNGAM for sample efficient LiNGAM recovery.

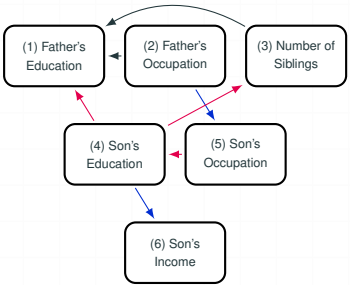
Graphs Estimated by OptLiNGAM and DirectLiNGAM



(a) True Graph



(b) OptLiNGAM



(c) DirectLiNGAM

- OptLiNGAM successfully identifies most causal relationships between variables, except for two reversed edges, (4, 1) and (6, 5).
- DirectLiNGAM fails to capture important connections, such as (2, 4) and (3, 5), and incorrectly assigns directions to edges, such as (4, 3) and (5, 4), all of which are correctly identified by OptLiNGAM.

Discussion

Summary

- Exogeneity identification optimized by reducing the conditioning set to the parent set.
- Sample efficient algorithm with sample complexity based on d_{in} without requiring incoherence or parental faithfulness.
- Sample optimality demonstrated by aligning lower and upper bounds of sample complexity, without prior knowledge of d_{in}.

Future works

- Computationally efficient algorithms while maintaining comparable sample efficiency.
- Consistency and potential optimality of the proposed algorithm under heavy-tailed errors.

Reference

- Shimizu, S. (2014). LiNGAM: Non-Gaussian methods for estimating causal structures. Behaviormetrika, 41(1), 65-98.
- Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., & Hoyer, P. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research-JMLR, 12(Apr), 1225-1248.
- Wang, Y. S., & Drton, M. (2020). High-dimensional causal discovery under non-Gaussianity. Biometrika, 107(1), 41-59.
- Zhao, R., He, X., & Wang, J. (2022). Learning linear non-Gaussian directed acyclic graph with diverging number of nodes. Journal of Machine Learning Research, 23(269), 1-34.
- Gao, M., Tai, W. M., & Aragam, B. (2022, May). Optimal estimation of Gaussian DAG models. In International Conference on Artificial Intelligence and Statistics (pp. 8738-8757). PMLR.
- Ghoshal, A., & Honorio, J. (2018, March). Learning linear structural equation models in polynomial time and sample complexity. In International Conference on Artificial Intelligence and Statistics (pp. 1466-1475). PMLR.
- Park, G., Moon, S. J., Park, S., & Jeon, J. J. (2021). Learning a high-dimensional linear structural equation model via l1-regularized regression. Journal of Machine Learning Research, 22(102), 1-41.
- Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Machine learning, 65, 31-78.
- Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). Socioeconomic background and achievement. Seminar Press.
- Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. Revue de l'Institut international de statistique, 2-8.
- Skitovitch, V. P. (1953). On a property of the normal distribution. DAN SSSR, 89:217–219.

Thank you!

Appendix - Darmois-Skitovitch Theorem

Darmois-Skitovitch Theorem (Darmois, 1953; Skitovitch, 1953)

Define two random variables u_1 and u_2 as linear combinations of independent random variables $s_i, i = 1, \ldots, m$, such that

$$u_1 = \sum_{i=1}^{m} c_{1,i} s_i \quad \text{and} \quad u_2 = \sum_{i=1}^{m} c_{2,i} s_i.$$

If u_1 and u_2 are independent, all variables s_i with $c_{1,i} c_{2,i} \neq 0$ are Gaussian distributed.

Appendix - Distance Covariance

$$T(e_{j,C}, X_k) = \frac{\mathrm{dcov}^2(e_{j,C}, X_k)}{I_{jk,2}}, \quad \text{and} \quad \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k) = \frac{\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_k)}{\widehat{I}_{jk,2}}.$$

Here, for any set $S \subset V \setminus \{j\}$, $\mathrm{dcov}^2(e_{j,C}, X_S) = I_{jS,1} + I_{jS,2} - 2I_{jS,3}$ with

$$\begin{aligned}
I_{jS,1} &= \mathbb{E}[|e_{j,C} - e'_{j,C}||X_S - X'_S||], \\
I_{jS,2} &= \mathbb{E}[|e_{j,C} - e'_{j,C}|]\mathbb{E}[||X_S - X'_S||], \\
I_{jS,3} &= \mathbb{E}\left[\mathbb{E}[|e_{j,C} - e'_{j,C}| \mid e_{j,C}]\mathbb{E}[||X_S - X'_S|| \mid X_S]\right],
\end{aligned}$$

where the notation $'$ denotes and independent copy of the corresponding random vector.
Additionally, $\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S) = \widehat{I}_{jS,1} + \widehat{I}_{jS,2} - 2\widehat{I}_{jS,3}$, where

$$\begin{aligned}
\widehat{I}_{jS,1} &= \frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}^{(i)}_{j,C} - \hat{e}^{(h)}_{j,C}||\mathbf{x}^{(i)}_S - \mathbf{x}^{(h)}_S||, \\
\widehat{I}_{jS,2} &= \left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}^{(i)}_{j,C} - \hat{e}^{(h)}_{j,C}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n} ||\mathbf{x}^{(i)}_S - \mathbf{x}^{(h)}_S||\right), \\
\widehat{I}_{jS,3} &= \frac{1}{n^3}\sum_{i,h,m=1}^{n} |\hat{e}^{(i)}_{j,C} - \hat{e}^{(m)}_{j,C}||\mathbf{x}^{(h)}_S - \mathbf{x}^{(m)}_S||.
\end{aligned}$$

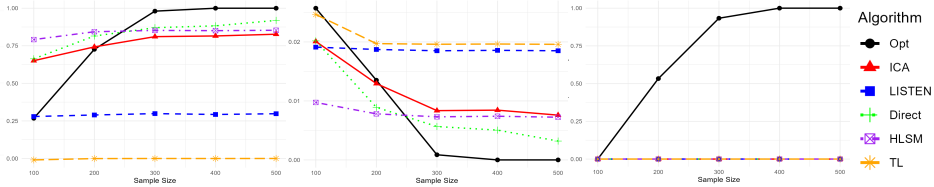
Appendix - Comparison to State-of-the-art Algorithms

Table 1: Empirical probability of successful graph recovery of all algorithms in high-dimensional settings.

(n, p)	**Method**	**Recovery Rate**	(n, p)	**Method**	**Recovery Rate**
(300,400)	Opt	**0.5333**	(300,500)	Opt	**0.3**
	TL	0		TL	0
	LISTEN	0		LISTEN	0
	HLSM	0		HLSM	0

- The proposed method achieves optimal sample complexity at a high computational cost; if graph recovery exceeds 24 hours, it is considered a failure.
- OptLiNGAM achieves empirical probabilities of successful graph recovery of 0.5333 and 0.3 for $p = 400$ and $p = 500$, respectively.
- None of the comparison methods successfully recover the true graph.

Appendix - Graph Structure Learning in Different Simulation Setting



(a) MCC (b) HM (c) Graph recovery rate

Figure 2: Average MCC, HM, and true graph recovery rate for 50-node graphs with $d_{in} = 1$ and $n \in \{100, 200, ..., 500\}$.

- Start with a three-node graph that has directed edges from node 1 to nodes 2 and 3.
- A new node is added at each step with a directed edge, where the probability of an existing node connecting to the new node depends on its number of neighbors.
- Only OptLiNGAM consistently achieves true graph recovery across all sample sizes considered.