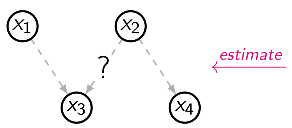
Differentiable Structure Learning and Causal Discovery for General Binary Data

Chang Deng, Bryon Aragam

Booth Business School, The University of Chicago

Background

• **Question**: Given data $\mathbf{X}$, how to learn a graph (DAG)?



estimate

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ 4.00 & -1.14 & 0.20 & -2.37 \\ -1.05 & 0.35 & -0.66 & -0.39 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

This is "Causal Discovery".

• Continuous DAG learning:

$$\min_{W \in \mathbb{R}^{p\times p}} s(W;\mathbf{X}) \quad \text{subject to} \quad h(W) = 0. \qquad (1)$$

Constraint: $h(W) = 0 \Leftrightarrow W$ is a DAG.

Related works

Consider **discrete** data $\mathbf{X}$, previous works:

$$\text{Data } \mathbf{X} \overset{causal\ disccovery}{\Longrightarrow} \text{Graph } G$$

(I) Require strong model assumptions (Linear, etc..)
(II) Misalign with dependence of general discrete data
(III) Mistreat the discrete value as continuous.
(IV) Lack general identifiability guarantee.

Contribution

Setting: for any binary data† $\mathbf{X}$ (no assumption).

Goal: recover G from $\mathbf{X}$.

- Prove DAGs are non-identifiable.
- Characterize all DAGs consistent with $\mathbf{X}$.
- Cast it as a continuous optimization problem.
- Prove Identifiability under weak conditions.
- Introduce BiNOTEARS, a general framework.

†: Extends to general discrete data with only notational changes, no new techniques.

General discrete data

- **Any** binary data matrix $\mathbf{X} \in \{0,1\}^{n\times p}$ can be modeled by *Multivariate Bernoulli Distribution (MVB)*.
- **MVB**: $X \sim \text{MultiBernoulli}(\boldsymbol{p})$ where $X = (X_1, \ldots, X_p) \in \{0,1\}^p$, its distribution follows

$$\mathbb{P}(X = x) = \prod_{S\subseteq[p]} \mathbb{P}(1_S)^{\prod_{j\in S} x_j \prod_{j\notin S}(1-x_j)} = \exp\left(\sum_{S\subseteq[p]} f^S B^S(x)\right)$$

$B^S(x) = \prod_{j\in S} x_j$. MVB characterizes **arbitrary** dependence within X.

- Conditional distribution includes higher order interaction:

$$\mathbb{P}(X_p = 1 \mid X_{-p}) = \sigma\left(\sum_{S\subseteq[p-1]} f^{S,p} B^S(x)\right) \qquad \sigma(z) = 1/(1+\exp(-z))$$

All the higher order interaction occurs for conditional distribution.

- $f^{S,p}$ can be recovered from logistic regression. The graph G can be recovered by

$$X_j \to X_j \Leftrightarrow \exists S \subseteq [p-1] \text{ with } j \in S, \text{ such that } f^{S,p} \neq 0 \Leftrightarrow \sum_{S\subseteq[p-1], j\in S} (f^{S,p})^2 > 0$$

- Simple example, $p = 3$.

$\mathbb{P}(X_3 = 1 \mid X_1, X_2) = \sigma\left(f^{\varnothing,3} + f^{1,3}X_1 + f^{2,3}X_2 + f^{12,3}X_1X_2\right)$. Then,

$$X_1 \to X_3 \Leftrightarrow (f^{1,3})^2 + (f^{12,3})^2 > 0, \quad X_2 \to X_3 \Leftrightarrow (f^{2,3})^2 + (f^{12,3})^2 > 0$$

Non-identifiability

Given any order π, write $\mathbb{P}(X) = \prod_{j=1}^{p} \mathbb{P}(X_{\pi(j)} \mid X_{\pi(1)}, \ldots, X_{\pi(j-1)})$.

So, for any observation $\mathbf{X} \in \{0,1\}^{n\times p}$:

Algorithm I

1. $\mathbf{X}_{\pi(j)} \overset{\text{logistic}}{\sim}$ all iteractions term of $(\mathbf{X}_{\pi(1)}, \ldots, \mathbf{X}_{\pi(j-1)})$, get $f_{\pi,j}$
2. Recover all edges from $(X_{\pi(1)}, \ldots, X_{\pi(j-1)})$ to $X_{\pi(j)}$ from $f_{\pi,j}$
3. Recover G_π from step 2.

Theorem (Informal): For $X \sim \text{MultiBernoulli}(\boldsymbol{p})$, under mild assumptions and as $n \to \infty$, **Algorithm I** returns, for any order π, an SEM $(f_{\pi,j}, G_\pi)$ that exactly recovers the distribution of X, and X is Markov w.r.t. G_π.

1. Any G_π is Markov to $\mathbb{P}(X)$, non-identifiability.
2. Algorithm I is purely combinatorial and fail to scale.

Continuous structure learning

$$\boldsymbol{H}_j = (\underbrace{h^{0,j}}_{\text{constant}}, \underbrace{h^{1,j}, \ldots, h^{p,j}}_{\text{first order}}, \underbrace{h^{12,j}, \ldots, h^{(p-1)p,j}}_{\text{second order}}, \underbrace{\cdots}_{\text{third to p th order}}, \underbrace{h^{123\cdots p,j}}_{\text{p-th order}})$$

Parameters: $\boldsymbol{H} = (\boldsymbol{H}_1, \ldots, \boldsymbol{H}_p) \in \mathbb{R}^{2^p\times p}$

Weighted adjacency matrix:

$$[W(\boldsymbol{H})]_{ij} = \sum_{S\subseteq[p], i\in S} \left(h^{S,j}\right)^2$$

Loss function: $\ell(\boldsymbol{H};\mathbf{X})$ [negative LL]

Regularizer (qausi-MCP):

$$\text{pen}_{\lambda,\delta}(t) = \lambda\left[\left(|t| - \frac{t^2}{2\delta}\right)\mathbf{1}(|t| < \delta) + \frac{\delta}{2}\mathbf{1}(|t| > \delta)\right]$$

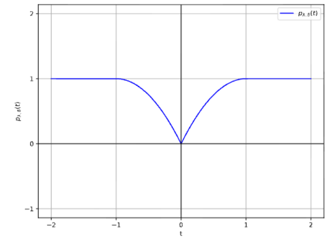


Figure 1: The plot of $p_{\lambda,\delta}(t)$ with $\lambda = 2, \delta = 1$

Score functions: $s(\boldsymbol{H};\lambda,\delta,\mathbf{X}) = \ell(\boldsymbol{H};\mathbf{X}) + \text{pen}_{\lambda,\delta}(W(\boldsymbol{H}))$

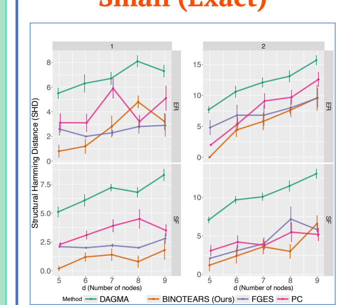
Forbidden self loop: $h^{S,j} = 0$, whenever $j \in S, \forall j \in [p], \forall S \subseteq [p]$

$$\begin{aligned} \min_{\mathbf{H}} \quad & s(\mathbf{H};\lambda,\delta,\mathbf{X}) \\ \text{subject to} \quad & h(W(\mathbf{H})) = 0, \\ & h^{S,j} = 0 \text{ if } j \in S \quad \forall j \in [p], \forall S \subseteq [p]. \end{aligned} \qquad (2)$$

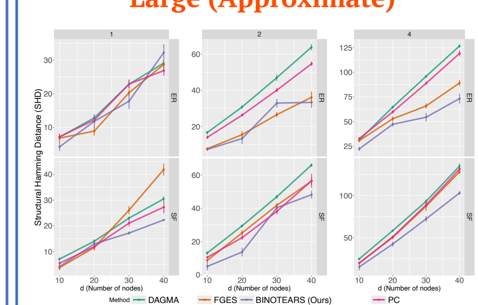
Theorem 2 (MEC): Let $X \sim \text{MultiBernoulli}(\boldsymbol{p})$. Under mild conditions and as $n \to \infty$, there exists small $\lambda, \delta > 0$ such that for any solutions $\boldsymbol{H}$ of (2) yields $W(\boldsymbol{H})$ in the same Markov Equivalence class, and $\boldsymbol{H}$ can be used to generate X exactly.

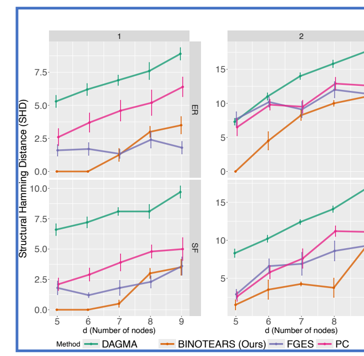
Experiment

Small (Exact)



Large (Approximate)





More in Paper!

arXiv: 2509.21658