Ordinal Causal Discovery

Yang Ni Bani Mallick

Department of Statistics Texas AM University

Presented by Wenbo Zhang

Table of Contents

Table of Contents

Causal Discovery

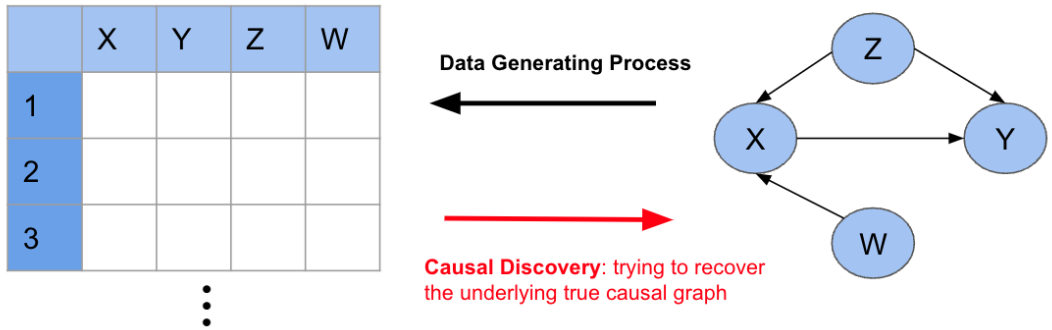
- Given a causal graph, we can do causal inference. What if we don't have the causal graph?

Causal Discovery

- Given a causal graph, we can do causal inference. What if we don't have the causal graph?
- Goal of Causal Discovery: identify the causal graph from data. (Here we only consider using observational data)

Causal Discovery

- Given a causal graph, we can do causal inference. What if we don't have the causal graph?
- Goal of Causal Discovery: identify the causal graph from data. (Here we only consider using observational data)
- Example:



X: Drug Dose
Y: Treatment Effects
Z: Age
W: Income

Causal Graph

Definition

Consider a set of random variables $\mathrm{X} = (X_1, \ldots, X_p)$ with index set $\mathrm{V} := \{1, \ldots, p\}$. A causal graph of X is a directed acyclic graph (DAG). Each node corresponds to a single variable X_i, where $i \in \{1, \ldots, p\}$ and each edge represents a direct causal relationship between variables.

Causal Graph

Definition

Consider a set of random variables $\mathrm{X} = (X_1, \ldots, X_p)$ with index set $\mathrm{V} := \{1, \ldots, p\}$. A causal graph of X is a directed acyclic graph (DAG). Each node corresponds to a single variable X_i, where $i \in \{1, \ldots, p\}$ and each edge represents a direct causal relationship between variables.

- Structural Equation Model (SEM)

$$X_j = f_j\left(\mathrm{PA}_j, N_j\right), \quad j = 1, \ldots, p$$

Definitions

Definition: d-separation

Two disjoint subsets of vertices A and B are d-separated by a third (also disjoint) subset C if every path between nodes in A and B is blocked by C.

Definition: Markov

The joint distribution $\mathcal{L}(\mathrm{X})$ is said to be Markov with respect to the DAG $\mathcal{G}$ if

$$\mathrm{A}, \mathrm{B} \quad d\text{-sep. by } \mathrm{C} \Rightarrow \mathrm{A} \perp \mathrm{B} \mid \mathrm{C}$$

Notation

We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov with respect to $\mathcal{G}$:

$$\mathcal{M}(\mathcal{G}) := \{\mathcal{L}(\mathrm{X}) : \mathcal{L}(\mathrm{X}) \text{ is Markov w.r.t. } \mathcal{G}\}$$

.

Markov Equivalence Class (MEC)

Definition: Markov Equivalent

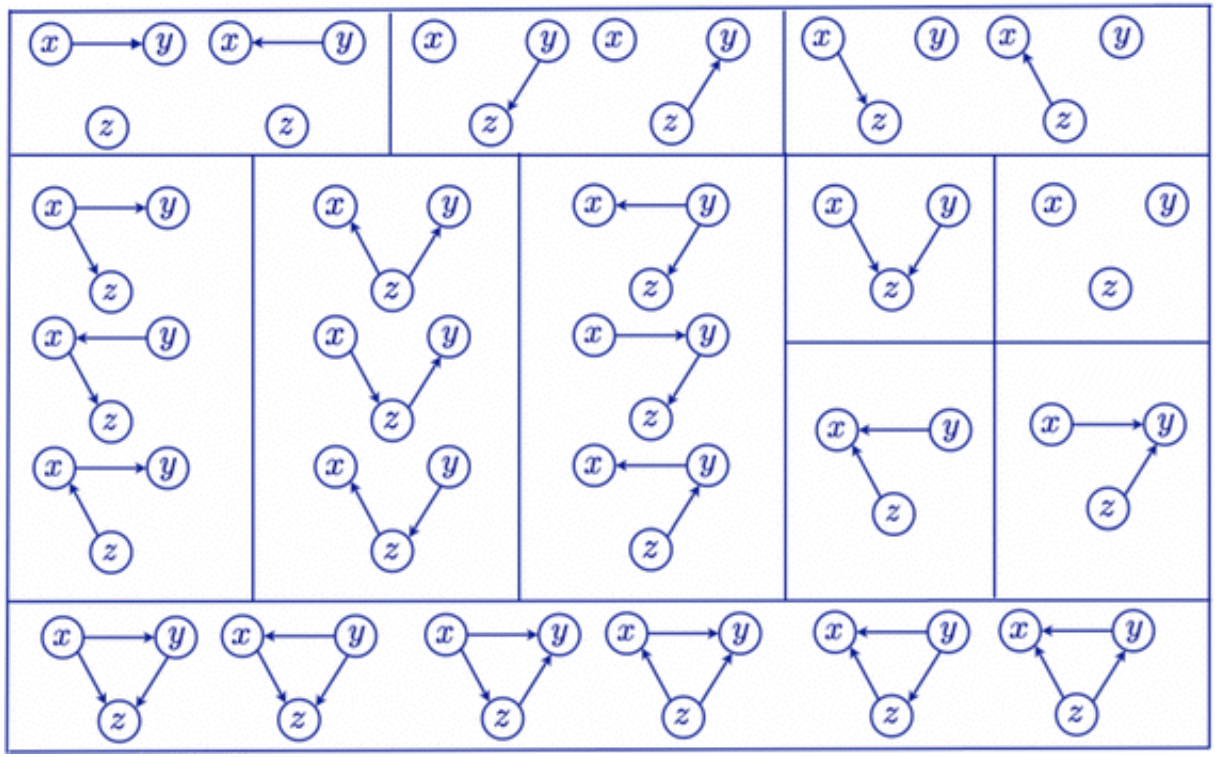
Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations, which means the Markov condition entails the same set of (conditional) independence conditions.

Definition: Markov Equivalence Class (MEC)

If the set of DAGs are Markov equivalent to some DAG (e.g., encode the same set of conditional independence conditions), then they are in the Markov equivalence class (MEC).

Example of MEC with 3 Nodes



From https://link.springer.com/article/10.1007/s41060-016-0038-6

Lemma

Two DAGs are Markov equivalent if and only if they have the same skeleton and the same immoralities.

Important Facts

Assumption: Faithfulness and Causal Su□ciency

- Faithfulness: A, B *d*-sep. by C $\Leftarrow$ A $\perp$ B | C
- Causal Sufficiency: There are no unobserved confounders of any of the variables in the graph.

Facts

Without loss of generality, given the above assumptions and no further parametric or semi-parametric assumptions, we can only identify the DAG up to its Markov equivalence class (MEC) from observational data,

But not all further assumptions are helpful. There are a few examples

- Linear Gaussian: only identify MEC
- Multinomial: only identify MEC
- Nonlinear Additive Noise: can identify true DAG ...

Table of Contents

Motivation

- While causal discovery for purely observational categorical data has been extensively studied, the vast majority of existing methods have exclusively focused on nominal (unordered) categorical variables.

Motivation

- While causal discovery for purely observational categorical data has been extensively studied, the vast majority of existing methods have exclusively focused on nominal (unordered) categorical variables.
- However, many categorical data contain ordinal information.

Motivation

- While causal discovery for purely observational categorical data has been extensively studied, the vast majority of existing methods have exclusively focused on nominal (unordered) categorical variables.
- However, many categorical data contain ordinal information.
- Example:
  psychologists often use questionnaires to measure latent traits such as personality and depression. The responses to those questionnaires are often categorical, say, with five levels (5-point Likert scale): "strongly disagree", "disagree", "neutral", "agree", and "strongly agree".

Motivation

- While causal discovery for purely observational categorical data has been extensively studied, the vast majority of existing methods have exclusively focused on nominal (unordered) categorical variables.
- However, many categorical data contain ordinal information.
- Example:
  psychologists often use questionnaires to measure latent traits such as personality and depression. The responses to those questionnaires are often categorical, say, with five levels (5-point Likert scale): "strongly disagree", "disagree", "neutral", "agree", and "strongly agree".
- Utilizing ordinal information of categorical data in causal discovery should be beneficial.

Table of Contents

Bivariate Method

- Firstly, an Ordinal Causal Discovery (OCD) method for bivariate data is introduced. Let $(X, Y) \in \{1, \ldots, S\} \times \{1, \ldots, L\}$ denote a pair of ordinal variables with S and L levels.

Bivariate Method

- Firstly, an Ordinal Causal Discovery (OCD) method for bivariate data is introduced. Let $(X, Y) \in \{1, \ldots, S\} \times \{1, \ldots, L\}$ denote a pair of ordinal variables with S and L levels.
- Model:
  The bivariate OCD considers the following probability distribution for causal model $X \to Y$,

$$p_{X \to Y}(X, Y) = p(X)p(Y \mid X),$$

  where $p(X)$is a multinomial/categorical distribution with probabilities $\pi = (\pi_1, \ldots, \pi_S)$ with $\sum_{s=1}^{S} \pi_s = 1$, and $p(Y \mid X)$ is defined by an ordinal regression model:

$$\Pr(Y \leq \ell \mid X = s) = F(\gamma_\ell - \beta_s), \quad \ell = 1, \ldots, L, \quad s = 1, \ldots, S$$

Bivariate Method

- Above equation implies the conditional probability distribution:

$$\Pr(Y = \ell \mid X = s) = F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s),$$

where $\gamma_0 = -\infty$ and $\gamma_L = \infty$.

Bivariate Method

- Above equation implies the conditional probability distribution:

$$\Pr(Y = \ell \mid X = s) = F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s),$$

  where $\gamma_0 = -\infty$ and $\gamma_L = \infty$.
- Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)$ and $\gamma = (\gamma_2, \ldots, \gamma_{L-1})$.
  We can denote the model $p_{X \to Y}$ by $p_{X \to Y}(X, Y \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
  Similarly, we denote the model $p_{Y \to X}$ by $p_{Y \to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$.

Bivariate Method

- Above equation implies the conditional probability distribution:

$$\Pr(Y = \ell \mid X = s) = F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s),$$

  where $\gamma_0 = -\infty$ and $\gamma_L = \infty$.
- Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)$ and $\gamma = (\gamma_2, \ldots, \gamma_{L-1})$.
  We can denote the model $p_{X \to Y}$ by $p_{X \to Y}(X, Y \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
  Similarly, we denote the model $p_{Y \to X}$ by $p_{Y \to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$.
- Method: given observations of (X, Y), we can get two maximal likelihood estimates (MLE) $\widehat{p}_{X \to Y}$ and $\widehat{p}_{Y \to X}$.
  If $\widehat{p}_{X \to Y} > \widehat{p}_{Y \to X}$, then $X \to Y$ is deemed a more likely data generating causal model.

Identifiability

Does the joint distribution $p(X, Y)$ tell us whether it has been induced by X to Y or from Y to X? In other words, is the structure identifiable from the joint distribution?

Definition: Distribution Equivalence

$p_{X\to Y}(X, Y \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $p_{Y\to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ are distribution equivalent if for any values of $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ there exist values of $(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ such that $p_{X\to Y}(X, Y \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = p_{Y\to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ for any X, Y, and vice versa.

- Distribution equivalent causal models are clearly not distinguishable from each other by examining their observational distributions.

Identifiability

- Example:
  The multinomial causal Bayesian Networks are distribution equivalent.

$P(Y\|X)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.16	0.84	0.16
$Y=2$	0.34	0.14	0.34
$Y=3$	0.50	0.02	0.50

(a)

X	$P(X)$
1	0.25
2	0.25
3	0.50

$P(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.040	0.210	0.080
$Y=2$	0.085	0.035	0.170
$Y=3$	0.125	0.005	0.250

(b)

$P(X\|Y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.12	0.29	0.33
$X=2$	0.64	0.12	0.01
$X=3$	0.24	0.59	0.66

(c)

Y	$P(Y)$
1	0.33
2	0.29
3	0.38

Identifiability

- Example:
  The multinomial causal Bayesian Networks are distribution equivalent.

$P(Y\|X)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.16	0.84	0.16
$Y=2$	0.34	0.14	0.34
$Y=3$	0.50	0.02	0.50

(a)

X	$P(X)$
1	0.25
2	0.25
3	0.50

$P(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.040	0.210	0.080
$Y=2$	0.085	0.035	0.170
$Y=3$	0.125	0.005	0.250

(b)

$P(X\|Y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.12	0.29	0.33
$X=2$	0.64	0.12	0.01
$X=3$	0.24	0.59	0.66

(c)

Y	$P(Y)$
1	0.33
2	0.29
3	0.38

- As the example shows we can find a set of parameters, i.e., the conditional $p(X \mid Y)$ and marginal $p(Y)$ probabilities of the reverse causal model $Y \to X$, which leads to the same joint distribution as the causal model. Therefore, not identifiable.

Identifiability

- Example:
  The multinomial causal Bayesian Networks are distribution equivalent.

$P(Y\|X)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.16	0.84	0.16
$Y=2$	0.34	0.14	0.34
$Y=3$	0.50	0.02	0.50

(a)

X	$P(X)$
1	0.25
2	0.25
3	0.50

$P(X\|Y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.12	0.29	0.33
$X=2$	0.64	0.12	0.01
$X=3$	0.24	0.59	0.66

(c)

Y	$P(Y)$
1	0.33
2	0.29
3	0.38

$P(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.040	0.210	0.080
$Y=2$	0.085	0.035	0.170
$Y=3$	0.125	0.005	0.250

(b)

- As the example shows we can find a set of parameters, i.e., the conditional $p(X \mid Y)$ and marginal $p(Y)$ probabilities of the reverse causal model $Y \to X$, which leads to the same joint distribution as the causal model. Therefore, not identifiable.
- However, if incorporating the ordinal information, we will show that $p_{X \to Y}(X, Y \mid \pi, \beta, \gamma)$ and $p_{Y \to X}(Y, X \mid \rho, \alpha, \eta)$ are generally not distribution equivalent and are, therefore, identifiable.

Identifiability

Theorem (Identifiability of OCD)

Let $X \in \{1, \ldots, S\}$ and $Y \in \{1, \ldots, L\}$ where $S, L \geq 2$ and $\max\{S, L\} \geq 3$. Suppose $X \to Y$ is the data generating causal model and the observational probability distribution of (X, Y) is given by:

$$p(X, Y) = p_{X \to Y}(X, Y \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

For almost all $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to the Lebesgue measure, the distribution cannot be equivalently represented by the reverse causal model, i.e., there does not exist $(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ such that,

$$p(X, Y) = p_{Y \to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}), \forall X, Y$$

Ordinal Bayesian Networks

- Given a large enough dataset, in the causal direction, $\widehat{p}_{X\to Y}(X, Y)$ can be arbitrarily close to the true $p(X, Y)$. However, there does not exist any set of parameter values in the reverse causal model $p_{Y\to X}(Y, X \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ that produces the conditional $p(X \mid Y)$ and marginal $p(Y)$ probability such that $\widehat{p}_{Y\to X}(X, Y)$ close $p(X, Y)$.

Ordinal Bayesian Networks

- Given a large enough dataset, in the causal direction, $\widehat{p}_{X\to Y}(X,Y)$ can be arbitrarily close to the true $p(X,Y)$. However, there does not exist any set of parameter values in the reverse causal model $p_{Y\to X}(Y,X \mid \boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\eta})$ that produces the conditional $p(X \mid Y)$ and marginal $p(Y)$ probability such that $\widehat{p}_{Y\to X}(X,Y)$ close $p(X,Y)$.
- Example: The ordinal causal Bayesian Networks

$P(Y\|X)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.16	0.84	0.16
$Y=2$	0.34	0.14	0.34
$Y=3$	0.50	0.02	0.50

(a)

×

X	$P(X)$
1	0.25
2	0.25
3	0.50

$P(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.040	0.210	0.080
$Y=2$	0.085	0.035	0.170
$Y=3$	0.125	0.005	0.250

(b)

$P(X\|Y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.12	0.29	0.33
$X=2$	0.64	0.12	0.01
$X=3$	0.24	0.59	0.66

(c)

×

Y	$P(Y)$
1	0.33
2	0.29
3	0.38

$\hat{P}(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.104	0.087	0.141
$Y=2$	0.065	0.071	0.154
$Y=3$	0.078	0.089	0.211

(d)

Ordinal Bayesian Networks

- Given a large enough dataset, in the causal direction, $\widehat{p}_{X\to Y}(X, Y)$ can be arbitrarily close to the true $p(X, Y)$. However, there does not exist any set of parameter values in the reverse causal model $p_{Y\to X}(Y, X \mid \rho, \alpha, \eta)$ that produces the conditional $p(X \mid Y)$ and marginal $p(Y)$ probability such that $\widehat{p}_{Y\to X}(X, Y)$ close $p(X, Y)$.
- Example: The ordinal causal Bayesian Networks

$P(Y\|X)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.16	0.84	0.16
$Y=2$	0.34	0.14	0.34
$Y=3$	0.50	0.02	0.50

(a)

×

X	$P(X)$
1	0.25
2	0.25
3	0.50

$P(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.040	0.210	0.080
$Y=2$	0.085	0.035	0.170
$Y=3$	0.125	0.005	0.250

(b)

$P(X\|Y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.12	0.29	0.33
$X=2$	0.64	0.12	0.01
$X=3$	0.24	0.59	0.66

(c)

×

Y	$P(Y)$
1	0.33
2	0.29
3	0.38

$\hat{P}(X,Y)$	$X=1$	$X=2$	$X=3$
$Y=1$	0.104	0.087	0.141
$Y=2$	0.065	0.071	0.154
$Y=3$	0.078	0.089	0.211

(d)

- Therefore, $p_{X\to Y}(X, Y \mid \pi, \beta, \gamma)$ can be distinguished from $p_{Y\to X}(Y, X \mid \rho, \alpha, \eta)$.

Multivariate Ordinal Causal Discovery

Let $\boldsymbol{X} = (X_1, \ldots, X_p) \in \{1, \ldots, L_1\} \times \cdots \times \{1, \ldots, L_p\}$ denote p ordinal variables. Let $G = (V, E)$ denote a causal BN with a set of nodes $V = \{1, \ldots, p\}$ directed edges E. Let $pa(j) = \{k \mid k \to j\} \subseteq V$ denote the set of parents of node j in G and let $\boldsymbol{X}_{pa(j)} = \{X_k \mid k \in pa(j)\}$. Given G, the joint distribution of $\boldsymbol{X}$ factorizes,

$$p(\boldsymbol{X} \mid G) = \prod_{j=1}^{p} p\left(X_j \mid \boldsymbol{X}_{pa(j)}\right),$$

where each conditional distribution $p\left(X_j \mid \boldsymbol{X}_{pa(j)}\right)$ is an ordinal regression model of which the cumulative distribution is given by, for $\ell = 1, \ldots, L_j$

$$\Pr\left(X_j \leq \ell \mid \boldsymbol{X}_{pa(j)}\right) = F\left(\gamma_{j\ell} - \sum_{k \in pa(j)} \beta_{jk} x_k - \alpha_j\right)$$

Multivariate Ordinal Causal Discovery

The implied conditional probability distribution

$$\begin{aligned}
&\Pr\left(X_j = \ell \mid \boldsymbol{X}_{pa(j)} = s\right) \\
&= F\left(\gamma_{j\ell} - \sum_{k \in pa(j)} \beta_{jkh_k} - \alpha_j\right) - F\left(\gamma_{j,\ell-1} - \sum_{k \in pa(j)} \beta_{jkh_k} - \alpha_j\right)
\end{aligned}$$

where $\ell = 1, \ldots, L_j$ and $\boldsymbol{s} \in \prod_{k \in \text{pa}(j)} \{1, \ldots, L_k\}$.

Causal Graph Structure Learning

A score-and-search learning algorithm is proposed to estimate the structure of causal graphs.

- Score. We score causal graphs by the Bayesian information criterion (BIC). Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denote n realizations of $\boldsymbol{X}$. The score of G (smaller is better) is given by

$$\mathrm{BIC}(G \mid \boldsymbol{x}) = -2 \sum_{i=1}^{n} \log \widehat{p}\,(\boldsymbol{x}_i \mid G) + K \log(n)$$

Causal Graph Structure Learning

A score-and-search learning algorithm is proposed to estimate the structure of causal graphs.

- Score. We score causal graphs by the Bayesian information criterion (BIC). Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denote n realizations of $\boldsymbol{X}$. The score of G (smaller is better) is given by

$$\mathrm{BIC}(G \mid \boldsymbol{x}) = -2 \sum_{i=1}^{n} \log \widehat{p}\,(\boldsymbol{x}_i \mid G) + K \log(n)$$

- Search. Exhaustive Search are utilized for small networks and Greedy Search is used for moderate-sized networks.

Greedy Search Algorithm

Algorithm 1 Greedy Search

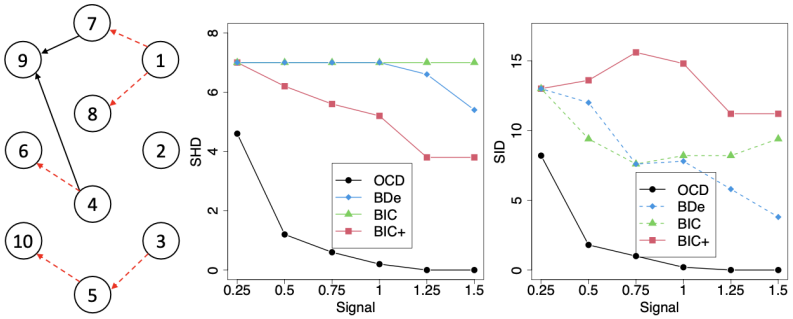
Input: data $\boldsymbol{x}$, initial graph G
Compute $\mathrm{BIC}(G|\boldsymbol{x})$ and set $\mathrm{BIC}_\star$=$\mathrm{BIC}(G|\boldsymbol{x})$.
repeat
  Initialize $Improvement = false$.
  for all graphs G' reachable from G **do**
    Compute $\mathrm{BIC}(G'|\boldsymbol{x})$.
    if $\mathrm{BIC}(G'|\boldsymbol{x}) < \mathrm{BIC}_\star$ **then**
      Set $G = G'$ and $\mathrm{BIC}_\star$=$\mathrm{BIC}(G'|\boldsymbol{x})$
      Set $Improvement = true$.
    end if
  end for
until $Improvement$ is $false$
Output: graph G

Table of Contents

Synthetic Data

Let $n = 500$ and the number of categories $L = 5$ for each node.

- Low-Dimensional Multivariate Ordinal Data

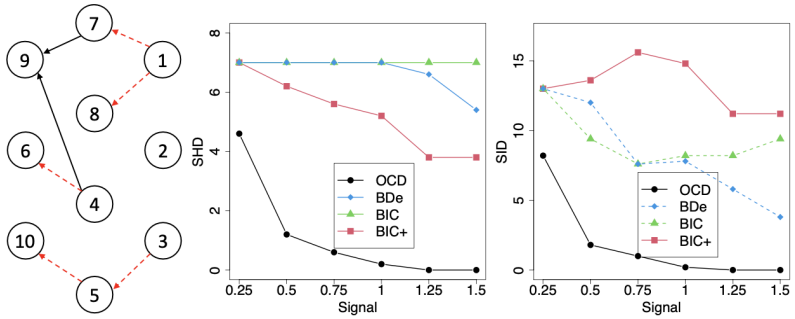


(a) True DAG (b) SHD (c) SID

Synthetic Data

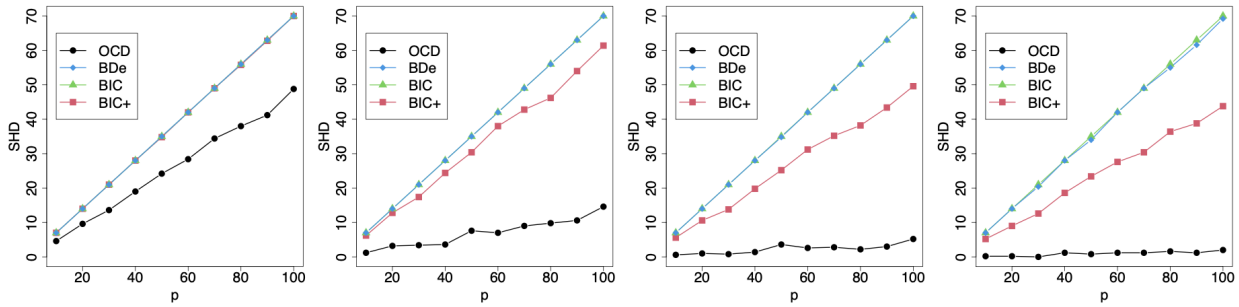
Let $n = 500$ and the number of categories $L = 5$ for each node.

- Low-Dimensional Multivariate Ordinal Data



(a) True DAG (b) SHD (c) SID

- High-Dimensional Multivariate Ordinal Data



(a) SHD in p ($\sigma = 0.25$) (b) SHD in p ($\sigma = 0.5$) (c) SHD in p ($\sigma = 0.75$) (d) SHD in p ($\sigma = 1$)

Real Data

- Sachs's Single-Cell Flow Cytometry Data
  853 cells and 11 phosphorylated proteins

	OCD	bQCD*	IGCI*	GR-AN*
SHD	14	15	15	16
SID	62	69	82	80
	HCR*	SLOPE*	ANM	LiNGAM
SHD	16	17	17	17
SID	76	86	78	86
	PC	CPC	GES	IAMB
SHD	18	18	18	20
SID	50-83	50-80	50-80	79-70
	BIC	BDe	MXM	RESIT
SHD	20	21	21	40
SID	53-77	49-104	49-104	45

Real Data

- Sachs's Single-Cell Flow Cytometry Data
  853 cells and 11 phosphorylated proteins

	OCD	bQCD*	IGCI*	GR-AN*
SHD	14	15	15	16
SID	62	69	82	80
	HCR*	SLOPE*	ANM	LiNGAM
SHD	16	17	17	17
SID	76	86	78	86
	PC	CPC	GES	IAMB
SHD	18	18	18	20
SID	50-83	50-80	50-80	79-70
	BIC	BDe	MXM	RESIT
SHD	20	21	21	40
SID	53-77	49-104	49-104	45

- Single-Cell RNA-Sequencing Data
  6701 pairs and 2 variables (transcription factor X and its target Y)

	OCD	HCR	bQCD	SLOPE
ACC	0.61	0.36	0.45	0.50
CPU	19m	22m	3.4h	2h

Homework

1. Explain why identifiability doesn't hold for Multinomial Bayesian Networks (Hint: the example from the paper I show).

Thank you!