Causal discovery: score-based and noise-based methods

Charles K. Assaad, Emilie Devijver, Eric Gaussier

charles.assaad@ens-lyon.fr

Table of content

Table of content

Recap about causal graphical models

Causal sufficiency $\forall X \leftarrow Z \rightarrow Y$, if $X, Y \in \mathcal{V}$ then $Z \in \mathcal{V}$.

Theorem (Markov equivalence for DAGs) Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent (have the same d-separations) *iff* they have the same skeleton and the same v-structures.

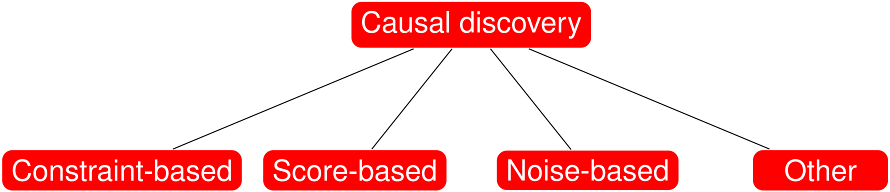
Completed partially directed acyclic graph (CPDAG) Let $[\mathcal{G}]$ be the Markov equivalence class of a DAG $\mathcal{G}$. The CPDAG $\mathcal{G}^*$ of $\mathcal{G}$ is the graph:

- With the same skeleton as $\mathcal{G}$;
- Where an edge is directed in $\mathcal{G}^*$ iff it occurs as a directed edge with the same orientation in every graph in $[\mathcal{G}]$;
- All other edges are undirected.

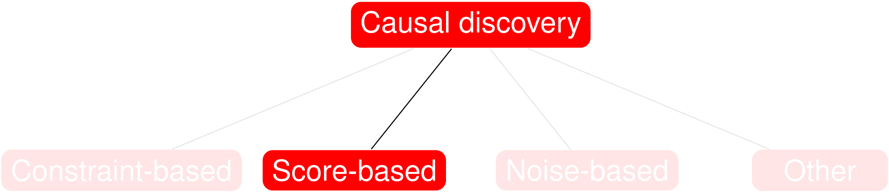
Faithfulness We say that a graph $\mathcal{G}$ and a compatible probability distribution P are faithful to one another if all and only the conditional independence relations true in P are entailed by the Markov condition applied to $\mathcal{G}$.

Table of content

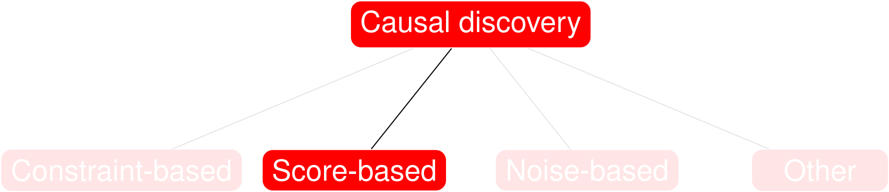
Causal discovery



Causal discovery



Causal discovery



Score-based: infer from observed data the equivalence class using a scoring criterion $S(\mathcal{G}, \mathbf{D})$.

Bayesian scoring criterion

Bayesian scoring criterion: $S_B(\mathcal{G}, \mathbf{D}) = \log P(\mathcal{G}) + \log P(\mathbf{D} \mid \mathcal{G})$

- $P(\mathcal{G})$: prior probability of $\mathcal{G}$
- $P(\mathbf{D} \mid \mathcal{G})$: marginal likelihood obtained by integrating over the unknown parameters the likelihood function applied to each observation in **D**

Bayesian information criterion (BIC - Schwarz, 1978) Under some assumptions:

$$S_B(\mathcal{G}, \mathbf{D}) = \underbrace{\log P(\mathbf{D} \mid \hat{\theta}, \mathcal{G}) - \frac{d}{2} \log m}_{BIC} + O(1)$$

$\hat{\theta}$: maximum-likelihood values of θ; d: number of free parameters; m: number of records in **D**; $O(1)$: constant

Bayesian scoring criterion

Bayesian scoring criterion: $S_B(\mathcal{G}, \mathbf{D}) = \log P(\mathcal{G}) + \log P(\mathbf{D} \mid \mathcal{G})$

- $P(\mathcal{G})$: prior probability of $\mathcal{G}$
- $P(\mathbf{D} \mid \mathcal{G})$: marginal likelihood obtained by integrating over the unknown parameters the likelihood function applied to each observation in **D**

Bayesian information criterion (BIC - Schwarz, 1978) Under some assumptions:

$$S_B(\mathcal{G}, \mathbf{D}) = \underbrace{\log P(\mathbf{D} \mid \hat{\theta}, \mathcal{G}) - \frac{d}{2} \log m}_{BIC} + O(1)$$

$\hat{\theta}$: maximum-likelihood values of θ; d: number of free parameters; m: number of records in **D**; $O(1)$: constant

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, Parents(X_i, \mathcal{G}))$

The Bayesian scoring criterion is decomposable

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \to X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, Parents(X_i, \mathcal{G}))$

The Bayesian scoring criterion is decomposable

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \rightarrow X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, Parents(X_i, \mathcal{G}))$

The Bayesian scoring criterion is decomposable

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \to X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, Parents(X_i, \mathcal{G}))$

The Bayesian scoring criterion is decomposable

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \rightarrow X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid Parents(X_j, \mathcal{G})$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Proposition If $\mathcal{G}$ and $\mathcal{G}'$ are in the same equivalence class, then $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{G}', \mathbf{D}) := S_B([\mathcal{G}], \mathbf{D})$

Proposition Let $[\mathcal{G}]$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of observations in $\mathbf{D}$. Then in the limit of large m, $S_B([\mathcal{G}], \mathbf{D}) > S_B([\mathcal{G}'], \mathbf{D})$ for $[\mathcal{G}] \neq [\mathcal{G}']$

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Proposition If $\mathcal{G}$ and $\mathcal{G}'$ are in the same equivalence class, then $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{G}', \mathbf{D}) := S_B([\mathcal{G}], \mathbf{D})$

Proposition Let $[\mathcal{G}]$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of observations in $\mathbf{D}$. Then in the limit of large m, $S_B([\mathcal{G}], \mathbf{D}) > S_B([\mathcal{G}'], \mathbf{D})$ for $[\mathcal{G}] \neq [\mathcal{G}']$

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Proposition If $\mathcal{G}$ and $\mathcal{G}'$ are in the same equivalence class, then $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{G}', \mathbf{D}) := S_B([\mathcal{G}], \mathbf{D})$

Proposition Let $[\mathcal{G}]$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of observations in $\mathbf{D}$. Then in the limit of large m, $S_B([\mathcal{G}], \mathbf{D}) > S_B([\mathcal{G}'], \mathbf{D})$ for $[\mathcal{G}] \neq [\mathcal{G}']$

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Proposition If $\mathcal{G}$ and $\mathcal{G}'$ are in the same equivalence class, then $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{G}', \mathbf{D}) := S_B([\mathcal{G}], \mathbf{D})$

Proposition Let $[\mathcal{G}]$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of observations in $\mathbf{D}$. Then in the limit of large m, $S_B([\mathcal{G}], \mathbf{D}) > S_B([\mathcal{G}'], \mathbf{D})$ for $[\mathcal{G}] \neq [\mathcal{G}']$

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Proposition If $\mathcal{G}$ and $\mathcal{G}'$ are in the same equivalence class, then $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{G}', \mathbf{D}) := S_B([\mathcal{G}], \mathbf{D})$

Proposition Let $[\mathcal{G}]$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of observations in **D**. Then in the limit of large m, $S_B([\mathcal{G}], \mathbf{D}) > S_B([\mathcal{G}'], \mathbf{D})$ for $[\mathcal{G}] \neq [\mathcal{G}']$

Neighbour classes

Covered edges An $X \to Y$ is covered in $\mathcal{G}$ if $Parents(Y, \mathcal{G}) = Parents(X, \mathcal{G}) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \to Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \to Y$ is covered in $\mathcal{G}$

Neighbour classes $[\mathcal{G}'] \in \mathcal{E}^+([\mathcal{G}])$ iff one can transform any DAG $\mathcal{G}$ in $[\mathcal{G}]$ to any DAG $\mathcal{G}'$ in $[\mathcal{G}']$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\mathcal{E}^-([\mathcal{G}])$ with a single edge deletion)

Neighbour classes

Covered edges An $X \to Y$ is covered in $\mathcal{G}$ if
$Parents(Y, \mathcal{G}) = Parents(X, \mathcal{G}) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \to Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \to Y$ is covered in $\mathcal{G}$

Neighbour classes $[\mathcal{G}'] \in \mathcal{E}^+([\mathcal{G}])$ iff one can transform any DAG $\mathcal{G}$ in $[\mathcal{G}]$ to any DAG $\mathcal{G}'$ in $[\mathcal{G}']$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\mathcal{E}^-([\mathcal{G}])$ with a single edge deletion)

Neighbour classes

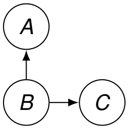
Covered edges An $X \to Y$ is covered in $\mathcal{G}$ if $Parents(Y, \mathcal{G}) = Parents(X, \mathcal{G}) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \to Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \to Y$ is covered in $\mathcal{G}$

Neighbour classes $[\mathcal{G}'] \in \mathcal{E}^+([\mathcal{G}])$ iff one can transform any DAG $\mathcal{G}$ in $[\mathcal{G}]$ to any DAG $\mathcal{G}'$ in $[\mathcal{G}']$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\mathcal{E}^-([\mathcal{G}])$ with a single edge deletion)

Example

What are the equivalence classes $[\mathcal{G}]$, $\mathcal{E}^{+}([\mathcal{G}])$ and $\mathcal{E}^{-}([\mathcal{G}])$ of the following graph $\mathcal{G}$?



GES: greedy equivalence search

GES algorithm

1. Initialisation: set $[\mathcal{G}]$ to the equivalence class corresponding to the DAG with no edge
2. Repeatedly replace $[\mathcal{G}]$ with the member of $\mathcal{E}^+([\mathcal{G}])$ that has the highest score, until no such replacement increases the score
3. Repeatedly replace $[\mathcal{G}]$ with the member of $\mathcal{E}^-([\mathcal{G}])$ that has the highest score, until no such replacement increases the score
4. Output the current class $[\mathcal{G}]$

Consistency of GES Let $[\mathcal{G}]$ denote the equivalence class that results from GES, let P denote a faithfull distribution of $\mathcal{G}$ associated with **D**, and let m denote the number of observations in **D**. Then in the limit of large m, $[\mathcal{G}]$ is a perfect map of P.

GES: greedy equivalence search

GES algorithm

1. Initialisation: set $[\mathcal{G}]$ to the equivalence class corresponding to the DAG with no edge
2. Repeatedly replace $[\mathcal{G}]$ with the member of $\mathcal{E}^+([\mathcal{G}])$ that has the highest score, until no such replacement increases the score
3. Repeatedly replace $[\mathcal{G}]$ with the member of $\mathcal{E}^-([\mathcal{G}])$ that has the highest score, until no such replacement increases the score
4. Output the current class $[\mathcal{G}]$

Consistency of GES Let $[\mathcal{G}]$ denote the equivalence class that results from GES, let P denote a faithfull distribution of $\mathcal{G}$ associated with **D**, and let m denote the number of observations in **D**. Then in the limit of large m, $[\mathcal{G}]$ is a perfect map of P.

Advantages and drawbacks

- Advantages:
  - Can discover the Markov equivalence class
- Drawbacks:
  - Can only discover the Markov equivalence class (or CPDAG);
  - High computational complexity (NP-hard):

d	Nombre de graphe pour d variables
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702929343
9	1213442454842881
10	4175098976430598143
15	2377252655534103549218021828637671925350 5

Advantages and drawabacks

- Advantages:
  - Can discover the Markov equivalence class
- Drawbacks:
  - Can only discover the Markov equivalence class (or CPDAG);
  - High computational complexity (NP-hard):

d	Nombre de graphe pour d variables
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702929343
9	1213442454842881
10	4175098976430598143
15	2377252655534103549218021828637671925350 5

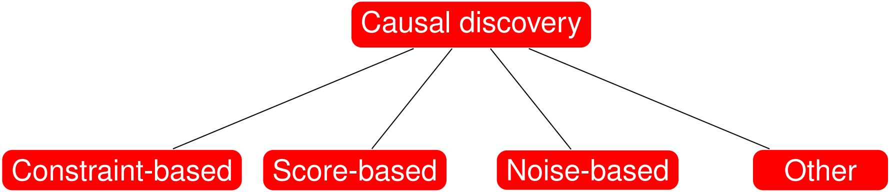
Extensions

1. Another (faster) approach exists based on the EM (expectation-maximisation) algorithm called MS-EM for *model selection EM* described in (Friedman, 1997)

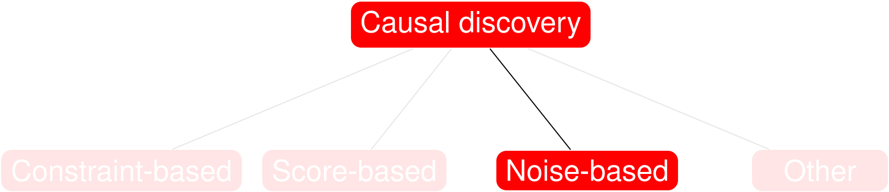
2. Several other extensions for different data types, *e.g.* for time series (Assaad *et al.*, 2022)

Table of content

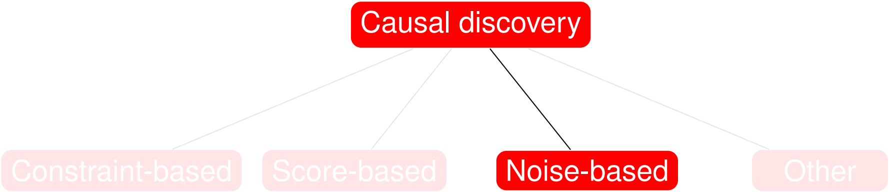
Causal discovery



Causal discovery



Causal discovery



Noise-based: find footprints in the noise that imply causal asymmetry.

Recap about causal graphical models

Topological ordering: Consider a causal DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a topological ordering $\mathcal{T} = \{X_1, \cdots, X_p\}$. If $X_i \rightarrow X_j$ in $\mathcal{G}$ then $i < j$.

Recap about structural causal models (1/2)

$V = \{X_1, X_2, \ldots, X_n\}$ set of endogenous variables
$U = \{\xi_1, \xi_2, \ldots, \xi_n\}$ corresponding set of exogenous variables.

Suppose that each endogenous variable X_i is a function of its parents in V together with ξ_i:

$$X_i = f_i(Parents(X_i), \xi_i).$$

Graphical representation is including only the endogenous variables V, and we use $Parents(X_i)$ to denote the set of endogenous parents of X_i.

Recap about structural causal models (1/2)

$V = \{X_1, X_2, \ldots, X_n\}$ set of endogenous variables
$U = \{\xi_1, \xi_2, \ldots, \xi_n\}$ corresponding set of exogenous variables.

Suppose that each endogenous variable X_i is a function of its parents in V together with ξ_i:

$$X_i = f_i(Parents(X_i), \xi_i).$$

Graphical representation is including only the endogenous variables V, and we use $Parents(X_i)$ to denote the set of endogenous parents of X_i.

Recap about structural causal models (1/2)

$V = \{X_1, X_2, \ldots, X_n\}$ set of endogenous variables
$U = \{\xi_1, \xi_2, \ldots, \xi_n\}$ corresponding set of exogenous variables.

Suppose that each endogenous variable X_i is a function of its parents in V together with ξ_i:

$$X_i = f_i(Parents(X_i), \xi_i).$$

Graphical representation is including only the endogenous variables V, and we use $Parents(X_i)$ to denote the set of endogenous parents of X_i.

Recap about structural causal models (2/2)

Independent Mechanism Principle
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions.

- Independence of noises, conditional independence of structures
- Independence of information contained in mechanisms
- Intervenability, autonomy, modularity, invariance, transfer

If the system of equations is acyclic, an assignment of values to the exogenous variables $\xi_1, \xi_2, \ldots, \xi_n$ uniquely determines the values of all the variables in the model. Then, if we have a probability distribution P' over the values of variables in ξ, this will induce a unique probability distribution P on V.

Recap about structural causal models (2/2)

Independent Mechanism Principle
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions.

- Independence of noises, conditional independence of structures
- Independence of information contained in mechanisms
- Intervenability, autonomy, modularity, invariance, transfer

If the system of equations is acyclic, an assignment of values to the exogenous variables $\xi_1, \xi_2, \ldots, \xi_n$ uniquely determines the values of all the variables in the model. Then, if we have a probability distribution P' over the values of variables in ξ , this will induce a unique probability distribution P on V.

The intuition behind the noise (1/2)

$$Suppose \begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$$

The intuition behind the noise (1/2)

$$\textit{Suppose}\begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$$

Given $P(X, Y)$, one can detect $X - Y$ but what about orientation?

The intuition behind the noise (1/2)

$$\text{Suppose}\begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$$

Given $P(X, Y)$, one can detect $X - Y$ but what about orientation?

$Y := 2X + \xi_y$?
or
$X := \frac{Y}{2} + \xi_x$?

The intuition behind the noise (1/2)

Suppose $\begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$

Given $P(X, Y)$, one can detect $X - Y$ but what about orientation?

$Y := 2X + \xi_y$?
or
$X := \frac{Y}{2} + \xi_x$?

Wihout further assumption we cannot know.

The intuition behind the noise (1/2)

$$\textit{Suppose} \begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$$

Given $P(X, Y)$, one can detect $X - Y$ but what about orientation?

$Y := 2X + \xi_y$?
or
$X := \frac{Y}{2} + \xi_x$?

Wihout further assumption we cannot know.

Assume that the noise follow a uniform distribution on $\{-1, 0, 1\}$

The intuition behind the noise (1/2)

Suppose $\begin{cases} X := \xi_x \\ Y := 2X + \xi_y \end{cases}$

Given $P(X, Y)$, one can detect $X - Y$ but what about orientation?

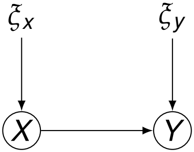
$Y := 2X + \xi_y$?
or
$X := \frac{Y}{2} + \xi_x$?

Wihout further assumption we cannot know.

Assume that the noise follow a uniform distribution on $\{-1, 0, 1\}$

X	Y	$\xi_y = Y - 2X$	$\xi_x = X - Y/2$
1	2	$0 \in \{-1, 0, 1\}$	$0 \in \{-1, 0, 1\}$
3	6	$0 \in \{-1, 0, 1\}$	$0 \in \{-1, 0, 1\}$
4	9	$1 \in \{-1, 0, 1\}$	$-0.5 \notin \{-1, 0, 1\}$

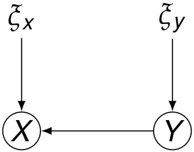
The intuition behind the noise (2/2)



$$M_1 : \begin{cases} X := f_x(\xi_x) \\ Y := f_y(X, \xi_y) \end{cases}$$

- $X \perp\!\!\!\perp_G \xi_y$
- $Y \not\!\perp\!\!\!\perp_G \xi_x$

Backwards model:



$$M_2 : \begin{cases} Y := g_y(\xi_y) \\ X := g_x(Y, \xi_x) \end{cases}$$

- $X \not\!\perp\!\!\!\perp_G \xi_y$
- $Y \perp\!\!\!\perp_G \xi_x$

Noise based question

Main question: Given $P(\mathcal{V})$ a compatible probability distribution of $\mathcal{G}$, can we discover $\mathcal{G}$?

Noise based question

Main question: Given $P(\mathcal{V})$ a compatible probability distribution of $\mathcal{G}$, can we discover $\mathcal{G}$? No!

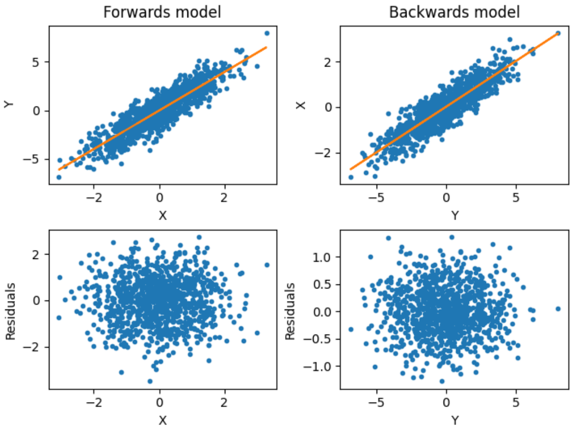
Noise based question

Main question: Given $P(\mathcal{V})$ a compatible probability distribution of $\mathcal{G}$, can we discover $\mathcal{G}$? No!
It is possible that $Y \perp\!\!\!\perp_P \xi_x$.

Noise based question

Main question: Given $P(\mathcal{V})$ a compatible probability distribution of $\mathcal{G}$, can we discover $\mathcal{G}$? No!
It is possible that $Y \perp\!\!\!\perp_P \xi_x$.
Example:

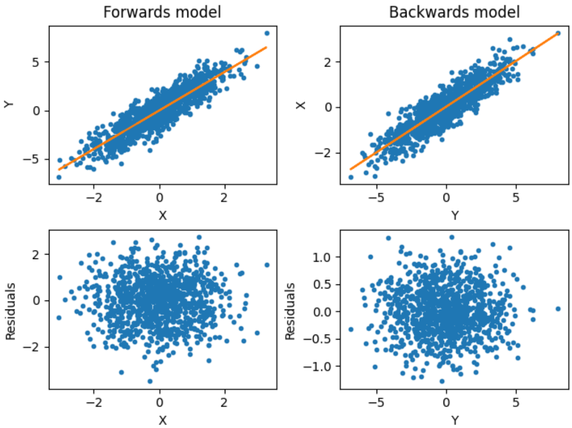
$$X \sim N(0,1)$$
$$\xi_y \sim N(0,1)$$
$$Y := 2X + \xi_y$$



Noise based question

Main question: Given $P(\mathcal{V})$ a compatible probability distribution of $\mathcal{G}$, can we discover $\mathcal{G}$? No!
It is possible that $Y \perp\!\!\!\perp_P \xi_x$.
Example:

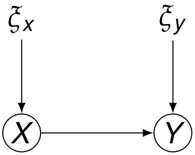
$X \sim N(0,1)$
$\xi_y \sim N(0,1)$
$Y := 2X + \xi_y$



$\Longrightarrow$ The Markov equivalence class is the best we can do!

Table of content

The linear case (1/2)

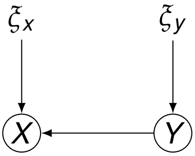


$$M_1 : \begin{cases} X := \xi_x \\ Y := aX + \xi_y \end{cases}$$

- $X \perp\!\!\!\perp_G \xi_y$
- $Y \not\perp\!\!\!\perp_G \xi_x$

When $Y \perp\!\!\!\perp_P \xi_x$?

Backwards model:



$$M_2 : \begin{cases} Y := \xi_y \\ X := bY + \xi_x \end{cases}$$

$$\begin{aligned} \xi_x &= X - bY \\ &= X - b(aX + \xi_y) \\ &= (1 - ba)X - b\xi_y \end{aligned}$$

The linear case (2/2)

$$Y = aX + \xi_y$$
$$\xi_x = (1 - ba)X - b\xi_y$$

When $Y \perp\!\!\!\perp_P \xi_x$?

The linear case (2/2)

$$Y = aX + \xi_y$$
$$\xi_x = (1 - ba)X - b\xi_y$$

When $Y \perp\!\!\!\perp_P \xi_x$?

Theorem (Darmois-Skitovich): Let $X_1, \cdots, X_n$ be independent, non degenerate random variables. If for two linear combinations:

$$l_1 = a_1 X_1 + \cdots + a_n X_n$$
$$l_2 = b_1 X_1 + \cdots + b_n X_n$$

are independent, then each X_i is normally distributed.

The linear non gaussian case (1/2)

Theorem (identiability of linear non-Gaussian models): Assume that $P(X, Y)$ admits the linear model

$$Y := aX + \xi_y, \qquad X \perp\!\!\!\perp_P \xi_y,$$

with continuous random variables X, ξ_y, and Y. Then there exists $b \in \mathbb{R}$ and a random variable ξ_x such that

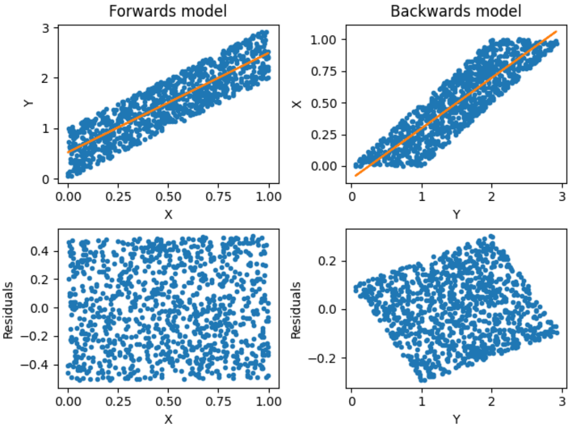
$$X := bY + \xi_x, \qquad Y \perp\!\!\!\perp_P \xi_x,$$

if and only if ξ_y and X are Gaussian.
(proof on board)

The linear non gaussian case (2/2)

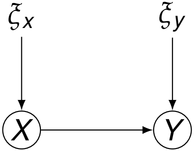
Example:

$$X \sim U(0, 1)$$
$$\xi_y \sim U(0, 1)$$
$$Y := 2X + \xi_y$$



The non linear case (1/3)

Continuous additive noise models



$$M_1 : \begin{cases} X := \xi_x \\ Y := f_y(X) + \xi_y \end{cases}$$

- $X \perp\!\!\!\perp_G \xi_y$
- $Y \not\perp\!\!\!\perp_G \xi_x$

When $Y \perp\!\!\!\perp_P \xi_x$?

The non linear case (2/3)

Theorem (identiability of additive noise models): Assume that $P(X, Y)$ admits the non-linear additive noise model

$$Y := f_y(X) + \xi_y, \qquad X \perp\!\!\!\perp_P \xi_y,$$

with continuous random variables X, ξ_y, and Y. Then there exists $g()$ and random variable ξ_x such that

$$X := g_x(Y) + \xi_x, \qquad Y \perp\!\!\!\perp_P \xi_x,$$

if and only if *Complicated Condition* is satisfied.
(Hoyer et al, 2008)

The non linear case (2/3)

Theorem (identiability of additive noise models): Assume that $P(X, Y)$ admits the non-linear additive noise model

$$Y := f_y(X) + \xi_y, \qquad X \perp\!\!\!\perp_P \xi_y,$$

with continuous random variables X, ξ_y, and Y. Then there exists $g()$ and random variable ξ_x such that

$$X := g_x(Y) + \xi_x, \qquad Y \perp\!\!\!\perp_P \xi_x,$$

if and only if *Complicated Condition* is satisfied.
(Hoyer et al, 2008)

Complicated Condition: The triple $(f_y, P(X), P(\xi_y))$ solves the following differential equation for all x, y with $(\log P(\xi_y))''(y - f_y(x))f'(x) \neq 0$.

The non linear case (3/3)

- The space that satisfy the condition is a 3-dimentional space;
  The space of continuous distributions is infinite dimensional;
  $\Longrightarrow$ we have identifiability for most distributions.
- If the noise is Gaussian, then the only functional form that satisfies Complicated Condition is linearity.
- If the function is linear and the noise is non-Gaussian, then one can't fit a linear backwards model **but** one can fit a non-linear backwards models.

Causal order discovery procedure in the bivariate case

Given $P(X, Y)$ and a dependence estimator $\hat{I}$

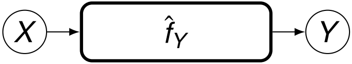
Procedure:

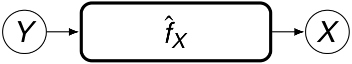
Causal order discovery procedure in the bivariate case

Given $P(X, Y)$ and a dependence estimator $\hat{I}$

Procedure:

1. Fit $\hat{f}_Y$ and $\hat{f}_X$:



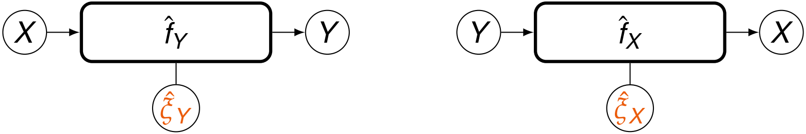


Given $P(X, Y)$ and a dependence estimator $\hat{I}$

Procedure:

1. Fit $\hat{f}_Y$ and $\hat{f}_X$:

2. Compute residuals $\hat{\xi}_Y$ and $\hat{\xi}_X$:



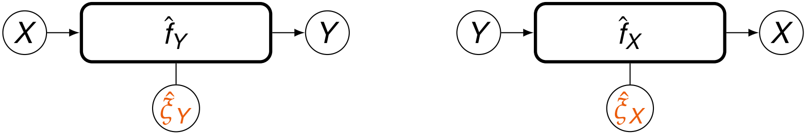
Causal order discovery procedure in the bivariate case

Given $P(X, Y)$ and a dependence estimator $\hat{I}$

Procedure:

1. Fit $\hat{f}_Y$ and $\hat{f}_X$:

2. Compute residuals $\hat{\xi}_Y$ and $\hat{\xi}_X$:



3. Order:
   - $\mathcal{T} = [X, Y]$ if $\hat{I}(x, \hat{\xi}_Y) < \hat{I}(y, \hat{\xi}_X)$
   - $\mathcal{T} = [Y, X]$ if $\hat{I}(y, \hat{\xi}_X) < \hat{I}(x, \hat{\xi}_Y)$

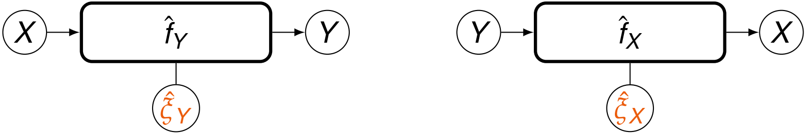
Causal order discovery procedure in the bivariate case

Given $P(X, Y)$ and a dependence estimator $\hat{I}$

Procedure:

1. Fit $\hat{f}_Y$ and $\hat{f}_X$:

2. Compute residuals $\hat{\xi}_Y$ and $\hat{\xi}_X$:



3. Order:
   - $\mathcal{T} = [X, Y]$ if $\hat{I}(x, \hat{\xi}_Y) < \hat{I}(y, \hat{\xi}_X)$
   - $\mathcal{T} = [Y, X]$ if $\hat{I}(y, \hat{\xi}_X) < \hat{I}(x, \hat{\xi}_Y)$
4. Output (suppose $\mathcal{T} = [X, Y]$):
   - $X \to Y$ if $X \perp\!\!\!\perp_P \hat{\xi}_Y$ and $Y \not\perp\!\!\!\perp_P \hat{\xi}_X$

Table of content

Minimality

Minimality condition A DAG $\mathcal{G}$ compatible with a probability distribution P is said to satisfy the minimality condition if P is not compatible with any proper subgraph of $\mathcal{G}$.

Minimality

Minimality condition A DAG $\mathcal{G}$ compatible with a probability distribution P is said to satisfy the minimality condition if P is not compatible with any proper subgraph of $\mathcal{G}$.

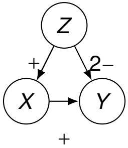
Remark: faithfulness $\Longrightarrow$ minimality.

Minimality and d-sep

Theorem (implication of minimality on d-sep): Consider the random vector $\mathcal{V}$ and assume that the joint distribution has a density with respect to a product measure. Suppose that $P(\mathcal{V})$ is Markov with respect to $\mathcal{G}$. Then $P(\mathcal{V})$ satisfies the minimality condition iff $\forall X \in \mathcal{V}$ and $\forall Y \in Parents(X, \mathcal{G})$, $X \not\perp\!\!\!\perp_P Y \mid Parents(X, \mathcal{G}) \setminus \{Y\}$.
(proof on board)

Violation of minimality

Example 1: canceling out



Example 2: constant functions

Linear non gaussian

Theorem (LiNGAM) Assume a linear SCM with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a compatible distribution $P(\mathcal{V})$ such that $\forall Y \in \mathcal{V}$

$$Y := \sum_{X \in Parents(Y, \mathcal{G})} a_{xy} X + \xi_y$$

where all ξ_y are jointly independent and non-Gaussian distributed. Additionally, we require that $\forall Y \in \mathcal{V}, X \in Parents(Y, \mathcal{G}), a_{xy} \neq 0$. Then, the graph $\mathcal{G}$ is identifiable from $P(\mathcal{V})$.
(proof in (Shimizu et al, 2011))

The LiNGAM algorithm

Algorithm 1 LiNGAM

Input: $P(\mathcal{V})$
Output: $\mathcal{G}$

```
Form an empty graph 𝒢 on vertex set 𝒱 = {X_1, ⋯, X_p}
Let S = {1, ⋯, p} and 𝒯 = []
repeat
    H = []
    for i ∈ S do
        for j ∈ S\{i} do
            ξ̂_ij = X_j − (cov(X_i, X_j) / var(X_i)) X_i
        end for
        h = Σ_{j∈S\{i}} Î(X_i, ξ̂_ij)
        H = [H, h]
    end for
    i* = arg min_{i∈S} H
    S = S\{i*}
    𝒯 = [𝒯, i*]
    ∀j ∈ S, X_j = ξ̂_{i*j}
until |S| = 0
Append(𝒯, S_0)
Construct a strictly lower triangular matrix by following the order in 𝒯, and estimate the connection strengths a_{i,j} by using some conventional covariance-based regression.
if a_{i,j} > 0 then
    Add X_i → X_j to 𝒢
end if
Return 𝒢
```

Where: $\hat{\xi}_{ij} = X_j - \frac{cov(X_i, X_j)}{var(X_i)} X_i$, $h = \sum_{j \in S \setminus \{i\}} \hat{I}(X_i, \hat{\xi}_{ij})$, $i^* = arg \min_{i \in S} H$, $\forall j \in S, X_j = \hat{\xi}_{i^* j}$.

Additive noise models

Theorem (ANM) Assume a non-linear SCM with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a compatible distribution $P(\mathcal{V})$ that satisfy the minimality condition with respect to $\mathcal{G}$. $\forall Y \in \mathcal{V}$

$$Y := f(Parents(Y, \mathcal{G})) + \xi_y$$

where all ξ_y are jointly independent. Then, the graph $\mathcal{G}$ is identifiable from $P(\mathcal{V})$.
(proof in (Peters et al, 2014))

Algorithm 2 ANM

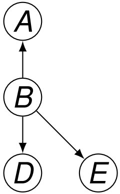
Input: $P(\mathcal{V})$

Output: $\mathcal{G}$

Form an empty graph $\mathcal{G}$ on vertex set $\mathcal{V} = \{X_1, \cdots, X_p\}$
Let $S = \{1, \cdots, p\}$ and $\mathcal{T} = []$
repeat
    $H = []$
    for $j \in S$ **do**
        $\hat{f}_{.j}$: Regress X^j on $\{X_i\}_{i \in S \setminus \{j\}}$
        $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$
        $h = \hat{I}(\{X_i\}_{i \in S \setminus \{j\}}, \xi_{.j})$
        $H = [H, h]$
    end for
    $i^* = arg \min_{i \in S} H$
    $S = S \setminus \{i^*\}$
    $\mathcal{T} = [i^*, \mathcal{T}]$
until $|S| = 0$
for $j \in \{2, \cdots, p\}$ **do**
    for $i \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\}$ **do**
        $\hat{f}_{.j}$: Regress X^j on $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \setminus \{i\}}$
        $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$
        if $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \setminus \{i\}} \not\perp\!\!\!\perp_P \xi_{.j}$ **then**
            Add $X_i \rightarrow X_j$ to $\mathcal{G}$
        end if
    end for
end for
Return $\mathcal{G}$

ANM in action (1/4)

- Suppose the true graph on right;
- Assumptions: CMC, minimality, causal sufficiency.



ANM in action (2/4)

- Estimate $A, B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{A, B, D\}, \hat{\xi}_e)$
- Estimate $A, D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{A, D, E\}, \hat{\xi}_b)$
- Estimate $A, B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{A, B, E\}, \hat{\xi}_d)$
- Estimate $B, D, E \mapsto A$ and $\hat{\xi}_a$
  - $H_4 = \hat{I}(\{B, D, E\}, \hat{\xi}_a)$

ANM in action (2/4)

- Estimate $A, B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{A, B, D\}, \hat{\xi}_e)$
- Estimate $A, D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{A, D, E\}, \hat{\xi}_b)$
- Estimate $A, B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{A, B, E\}, \hat{\xi}_d)$
- Estimate $B, D, E \mapsto A$ and $\hat{\xi}_a$
  - $H_4 = \hat{I}(\{B, D, E\}, \hat{\xi}_a)$

$$4 = Argmin(H)$$
$$\mathcal{T} = [A]$$

ANM in action (3/4)

- Estimate $B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{B, D\}, \hat{\xi}_e)$
- Estimate $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$

ANM in action (3/4)

- Estimate $B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{B, D\}, \hat{\xi}_e)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$
- Estimate $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$

$$1 = Argmin(H)$$
$$\mathcal{T} = [E, A]$$

ANM in action (3/4)

- Estimate $B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{B, D\}, \hat{\xi}_e)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$
- Estimate $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$

$$1 = Argmin(H)$$
$$\mathcal{T} = [E, A]$$

- Estimate $D \mapsto B$ and $\hat{\xi}_b$
  - $H_1 = \hat{I}(D, \hat{\xi}_b)$
- Estimate $B \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(B, \hat{\xi}_d)$

ANM in action (3/4)

- Estimate $B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{B, D\}, \hat{\xi}_e)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$
- Estimate $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$

$$1 = Argmin(H)$$
$$\mathcal{T} = [E, A]$$

- Estimate $D \mapsto B$ and $\hat{\xi}_b$
  - $H_1 = \hat{I}(D, \hat{\xi}_b)$
- Estimate $B \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(B, \hat{\xi}_d)$

$$2 = Argmin(H)$$
$$\mathcal{T} = [D, E, A]$$

ANM in action (3/4)

- Estimate $B, D \mapsto E$ and $\hat{\xi}_e$
  - $H_1 = \hat{I}(\{B, D\}, \hat{\xi}_e)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$
- Estimate $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$

$$1 = Argmin(H)$$
$$\mathcal{T} = [E, A]$$

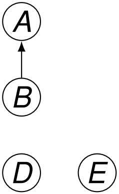
- Estimate $D \mapsto B$ and $\hat{\xi}_b$
  - $H_1 = \hat{I}(D, \hat{\xi}_b)$
- Estimate $B \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(B, \hat{\xi}_d)$

$$2 = Argmin(H)$$
$$\mathcal{T} = [D, E, A]$$

$$\mathcal{T} = [B, D, E, A]$$

ANM in action (4/4)

$$\mathcal{T} = [B, D, E, A]$$

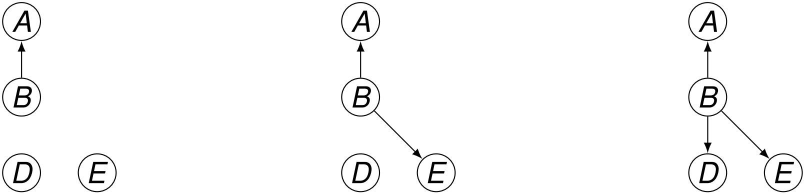


ANM in action (4/4)

$$\mathcal{T} = [B, D, E, A]$$

ANM in action (4/4)

$$\mathcal{T} = [B, D, E, A]$$



Advantages and drawabacks

- Advantages:
  - Can discovery the true graph;
  - Faithfulness is not needed.
- Drawbacks:
  - Semi parametric assumptions;
  - Need large sample size.

Advantages and drawabacks

- Advantages:
  - Can discovery the true graph;
  - Faithfulness is not needed.
- Drawbacks:
  - Semi parametric assumptions;
  - Need large sample size.

Some extensions

- Without causal sufficiency if linear relations;
- Extension to discrete additive noise models;
- Post non linear relations;
- Time series.

Some extensions

- Without causal sufficiency if linear relations;
- Extension to discrete additive noise models;
- Post non linear relations;
- Time series.

Some extensions

- Without causal sufficiency if linear relations;
- Extension to discrete additive noise models;
- Post non linear relations;
- Time series.

Some extensions

- Without causal sufficiency if linear relations;
- Extension to discrete additive noise models;
- Post non linear relations;
- Time series.

Exercise 1

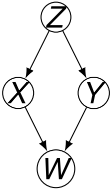
Why is faithfulness needed for constraint-based methods whereas noise-based methods only need minimality?

Exercise 2

After applying LiNGAM, how can you know if causal sufficiency is not respected?

Exercise 3

- Suppose the true graph on right;
- Assumptions: CMC, causal sufficiency, minimality;
- Generative process:

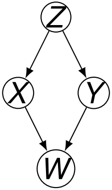


$$
\begin{aligned}
Z &= \xi_z & \xi_z &\sim U(0,1);\\
X &= a * Z + \xi_x & \xi_x &\sim U(0,1);\\
Y &= b * Z + \xi_y & \xi_y &\sim U(0,1);\\
W &= c * X - d * Y + \xi_w & \xi_w &\sim N(0,1).
\end{aligned}
$$

- Given a compatible distribution what would be the output of the LiNGAM algorithm? And what about the ANM algorithm?

Exercise 4

- Suppose the true graph on right;
- Assumptions: CMC, causal sufficiency, minimality;
- Generative process:



$$
\begin{aligned}
Z &= \xi_z & \xi_z &\sim U(0,1); \\
X &= Z^2 + \xi_x & \xi_x &\sim U(0,1); \\
Y &= Z^3 + \xi_y & \xi_y &\sim U(0,1); \\
W &= XY + \xi_w & \xi_w &\sim U(0,1).
\end{aligned}
$$

- Given a compatible distribution what would be the output of the LiNGAM algorithm? And what about the ANM algorithm?

Table of content

Conclusion

Score based:

- Under faithfulness, score-based methods can discover the Markov equivalence class (or CPDAG).

Noise based:

- Under linear non gaussian models noise-based methods can discover the causal graph;
- Under non-linear additive noise models noise-based methods can discover the causal graph.

Conclusion

Score based:

- Under faithfulness, score-based methods can discover the Markov equivalence class (or CPDAG).

Noise based:

- Under linear non gaussian models noise-based methods can discover the causal graph;
- Under non-linear additive noise models noise-based methods can discover the causal graph.

Conclusion

Score based:

- Under faithfulness, score-based methods can discover the Markov equivalence class (or CPDAG).

Noise based:

- Under linear non gaussian models noise-based methods can discover the causal graph;
- Under non-linear additive noise models noise-based methods can discover the causal graph.

References (1/3)

Direct inspirations (Part 1)

1. *Estimating the dimension of a model*, G. E. Schwarz, 1978
2. *A transformational characterization of Bayesian network structures*, D. M. Chickering, 1995
3. *Learning Bayesian networks is NP-complete*, D. M. Chickering, 1996
4. *Learning belief networks in the presence of missing values and hidden variables*, N. Friedman, 1997
5. *Optimal structure identification with greedy search*, D. M. Chickering, 2002

References (2/3)

Direct inspirations (Part 2)

1. *Elements of causal inference*, J. Peters, D. Janzing , B. Schölkopf. MIT Press, 2nd edition, 2017
2. *DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model*, S. Shimazu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. Hoyer, K. Bollen. JMLR, 2011
3. *Nonlinear causal discovery with additive noise models*, P. Hoyer, D. Janzing, J. Mooij, J. Peters, B. Schölkopf. Neurips, 2008
4. *Causal Discovery with Continuous Additive Noise Models*, J. Peters, J. Mooij, D. Janzing, B. Schölkopf. JMLR, 2014

References (3/3)

Additional readings

1. *Causal inference from noise*, N. Climenhaga, L. DesAutels, G. Ramsey. Noûs, 2019
2. *On the logic of causal models*, D. Geiger, J. Pearl. In Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, 1990
3. *A Linear Non-Gaussian Acyclic Model for Causal Discovery*, S. Shimazu, P. Hoyer, A. Hyvarinen, A. Kerminen. JMLR, 2006
4. *Analyse générale des liaisons stochastiques.*, G. Darmois. Review of the International Statistical Institute, 1953
5. *On a property of the normal distribution*, W. P. Skitovitch. Doklady Akademii Nauk SSSR, 89:217–219, 1953
6. *Causal Inference on Time Series using Restricted Structural Equation Models*, J. Peters, D. Janzing, B. Schölkopf. Neurips, 2013