Causal discovery: noise-based methods

Charles K. Assaad, Emilie Devijver, Eric Gaussier

emilie.devijver@univ-grenoble-alpes.fr

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

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, 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

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.

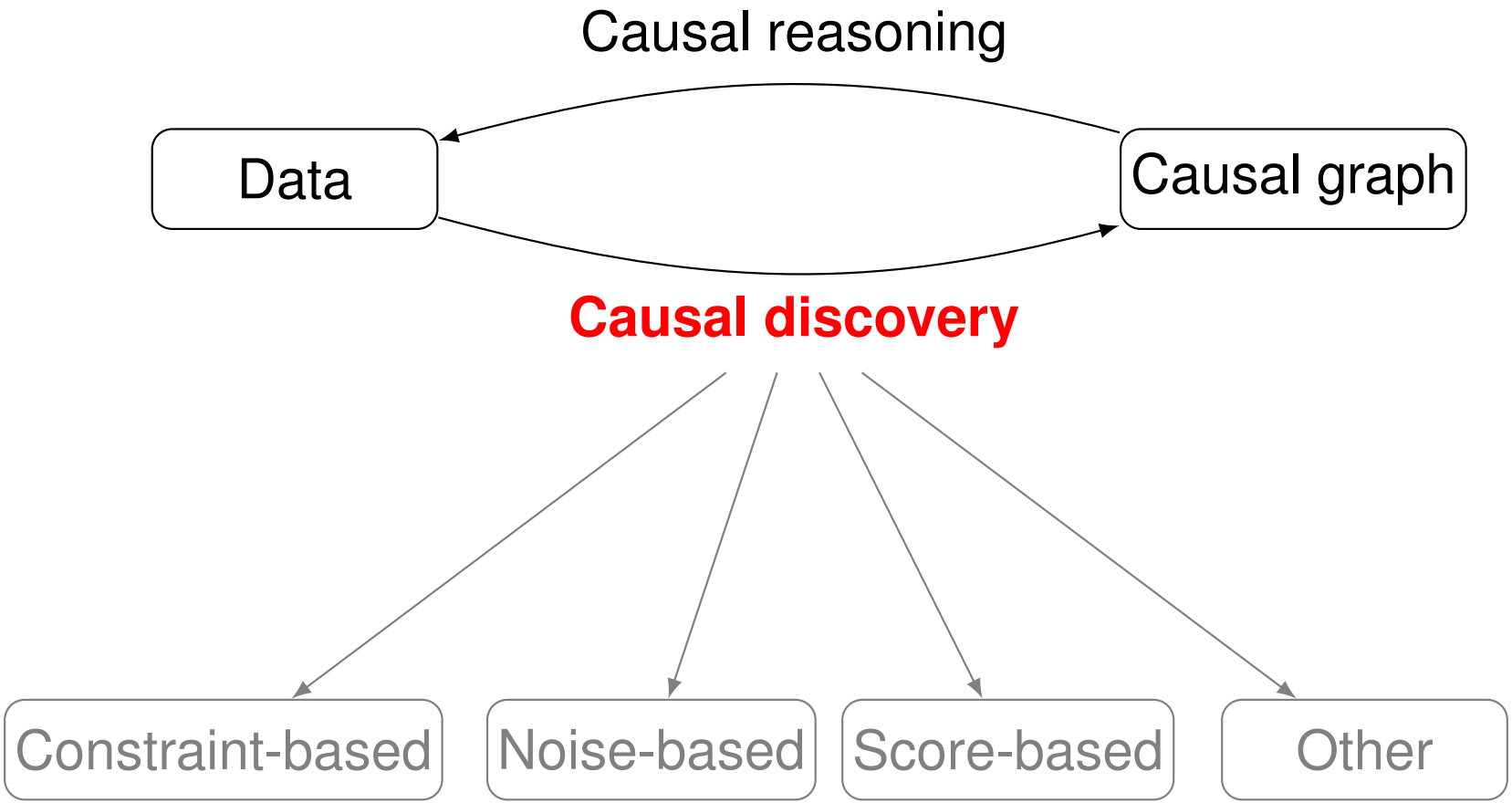
Recap about structural causal models (2/2)

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

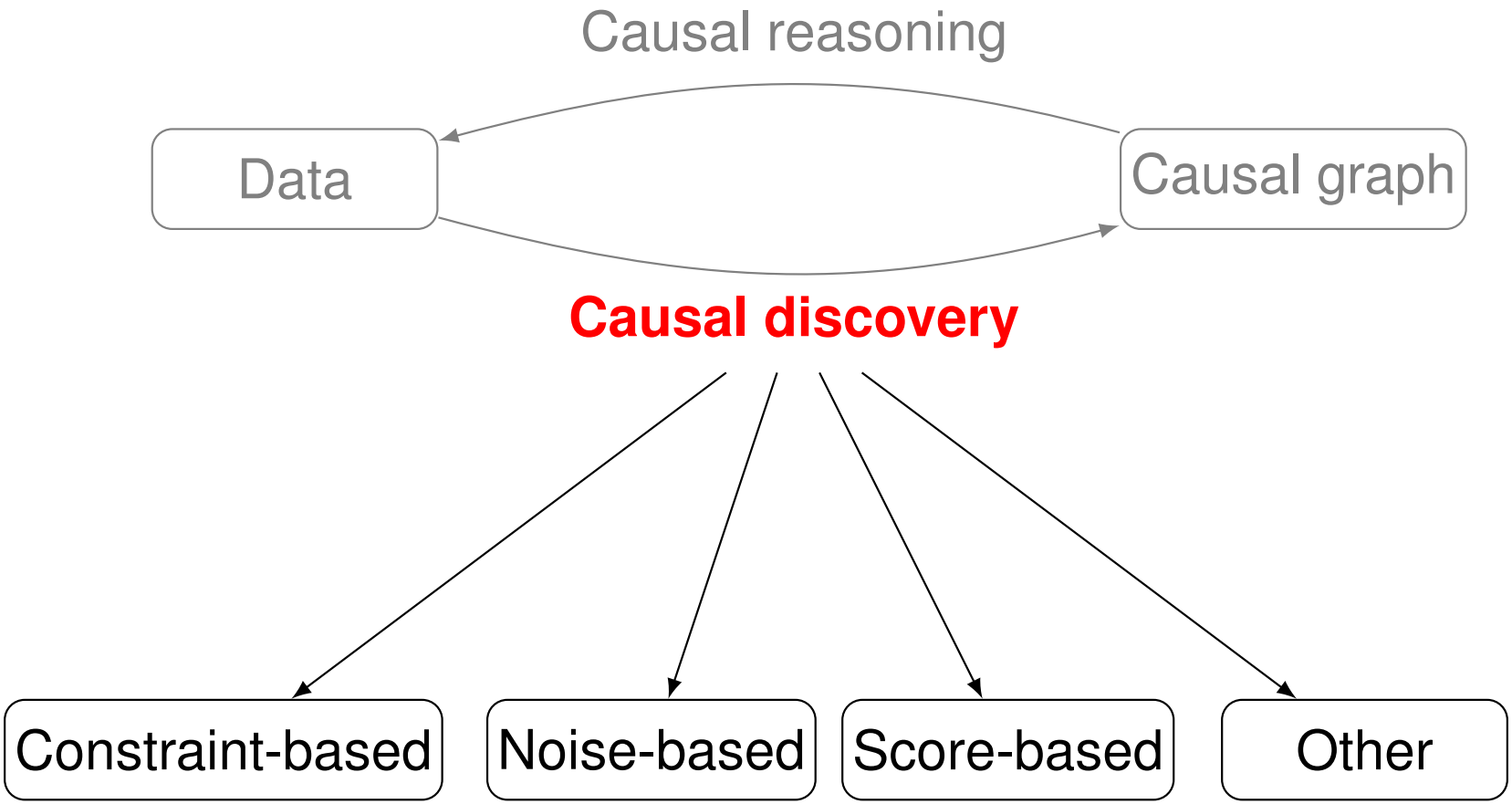
- Independence of noises, conditional independence of structures
- Independence of information contained in mechanisms

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.

Causal discovery



Causal discovery



Recap: Constraint-based

- Independence based algorithm to infer the causal graph
- Need to test independence → need of large sample size
- Faithfulness
- Identify the Markov equivalence class

Can we do better than identify the Markov equivalence class?
Not without parametric assumptions!

Recap: Constraint-based

- Independence based algorithm to infer the causal graph
- Need to test independence $\rightarrow$ need of large sample size
- Faithfulness
- Identify the Markov equivalence class

Can we do better than identify the Markov equivalence class?
Not without parametric assumptions!

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

Example with two nodes

Consider two variables X and Y, and assume we have infinite data and can deduce the joint distribution.

$$X \to Y \qquad \text{or} \qquad Y \to X$$

Markov equivalence class: $X - Y$

Example with two nodes

Consider two variables X and Y, and assume we have infinite data and can deduce the joint distribution.

$$X \to Y \qquad \text{or} \qquad Y \to X$$

Markov equivalence class: $X - Y$

Non-identifiability of two-node graphs

Proposition (Peters et al., 2017, p44)

For every joint distribution $P(x, y)$ on two real-valued random variables, there is an SCM in either direction that generates data consistent with $P(x, y)$.
Mathematically, there exists a function f_Y such that

$$Y = f_Y(X, U_Y), X \perp\!\!\!\perp U_Y$$

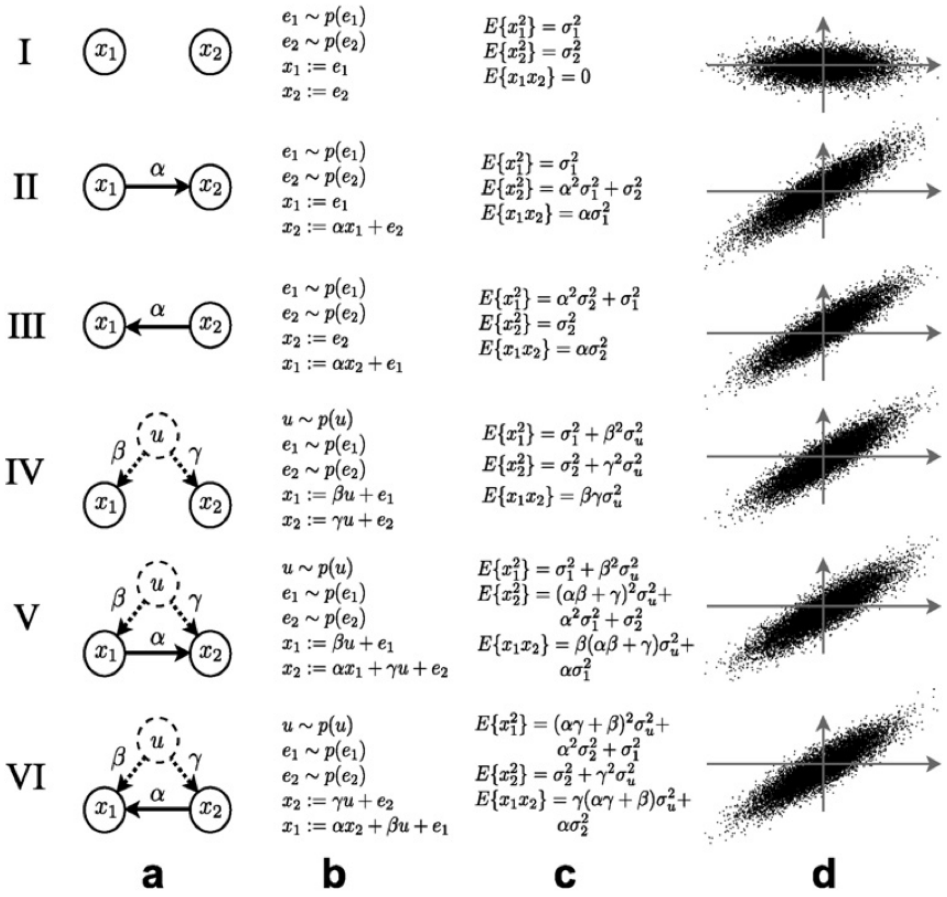
and there exists a function f_X such that

$$X = f_X(Y, U_X), Y \perp\!\!\!\perp U_X$$

where U_Y and U_X are real-valued random variables.

Non-identifiability

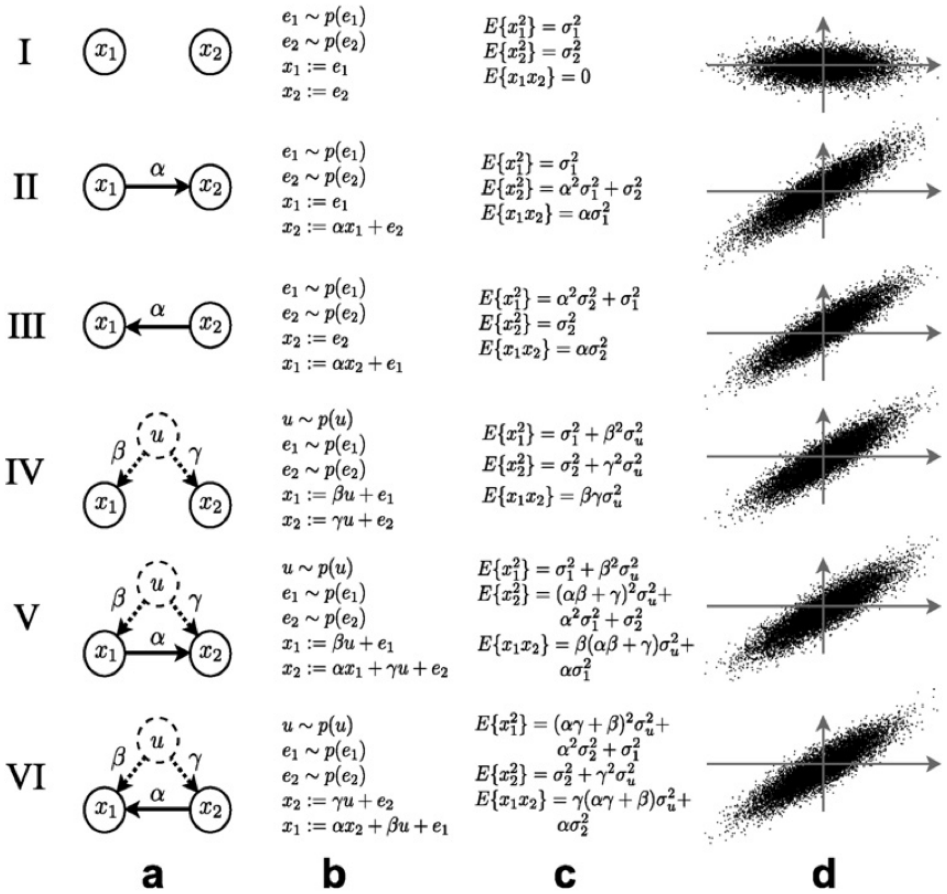
Ex from Hoyer et al. (2008)



The Markov equivalence class is the best we can do!

Non-identifiability

Ex from Hoyer et al. (2008)



The Markov equivalence class is the best we can do!

Non-identifiability

Multinomial or Gaussian distribution

Theorem (Markov Completeness, Meek (1995), Geiger and Pearl (1988))

If we have multinomial distributions or linear Gaussian structural equations, we can only identify a graph up to its Markov equivalence class.

The intuition behind the noise

$$\textit{Suppose} \begin{cases} X := \xi_X \\ Y := 2X + \xi_Y \end{cases}$$

The intuition behind the noise

$$\textit{Suppose} \begin{cases} X := \xi_X \\ Y := 2X + \xi_Y \end{cases}$$

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

The intuition behind the noise

$$\textit{Suppose} \begin{cases} X := \xi_X \\ Y := 2X + \xi_Y \end{cases}$$

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

$$Y := 2X + \xi_Y \qquad \text{or} \qquad X := \frac{Y}{2} + \xi_X ?$$

The intuition behind the noise

$$\textit{Suppose} \begin{cases} X := \xi_X \\ Y := 2X + \xi_Y \end{cases}$$

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

$$Y := 2X + \xi_Y \qquad \text{or} \qquad X := \frac{Y}{2} + \xi_X ?$$

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

The intuition behind the noise

$$\textit{Suppose} \begin{cases} X := \xi_X \\ Y := 2X + \xi_Y \end{cases}$$

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

$$Y := 2X + \xi_Y \qquad \text{or} \qquad X := \frac{Y}{2} + \xi_X ?$$

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	$1 \in \{-1, 0, 1\}$	$-0.5 \notin \{-1, 0, 1\}$

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

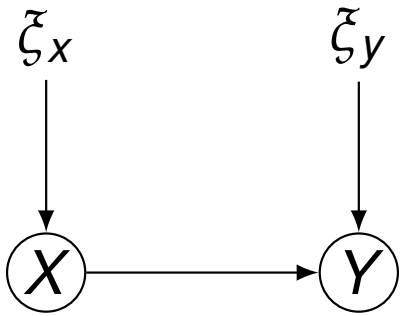
The post non linear case

In practice

Multivariate case

Conclusion

The linear case

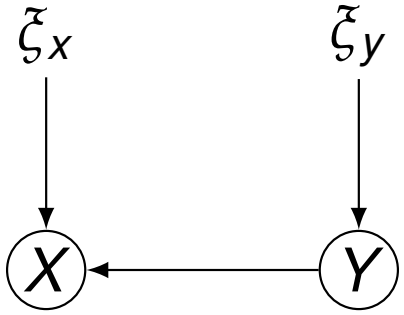


$$M_1 : \begin{cases} X := \xi_X \\ Y := aX + \xi_Y \end{cases}$$

- $X \perp\!\!\!\perp_G \xi_Y$
- $Y \not\perp\!\!\!\perp_G \xi_X$

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

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

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

Theorem (Darmois-Skitovich, 1953, 1954)

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 case

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

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

Theorem (Darmois-Skitovich, 1953, 1954)

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

Theorem (Identifiability 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.

It can be extended to multiple variables (see Shimizu et al. (2006), Peters et al. (2017)).

The linear non gaussian case

Theorem (Identifiability 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

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

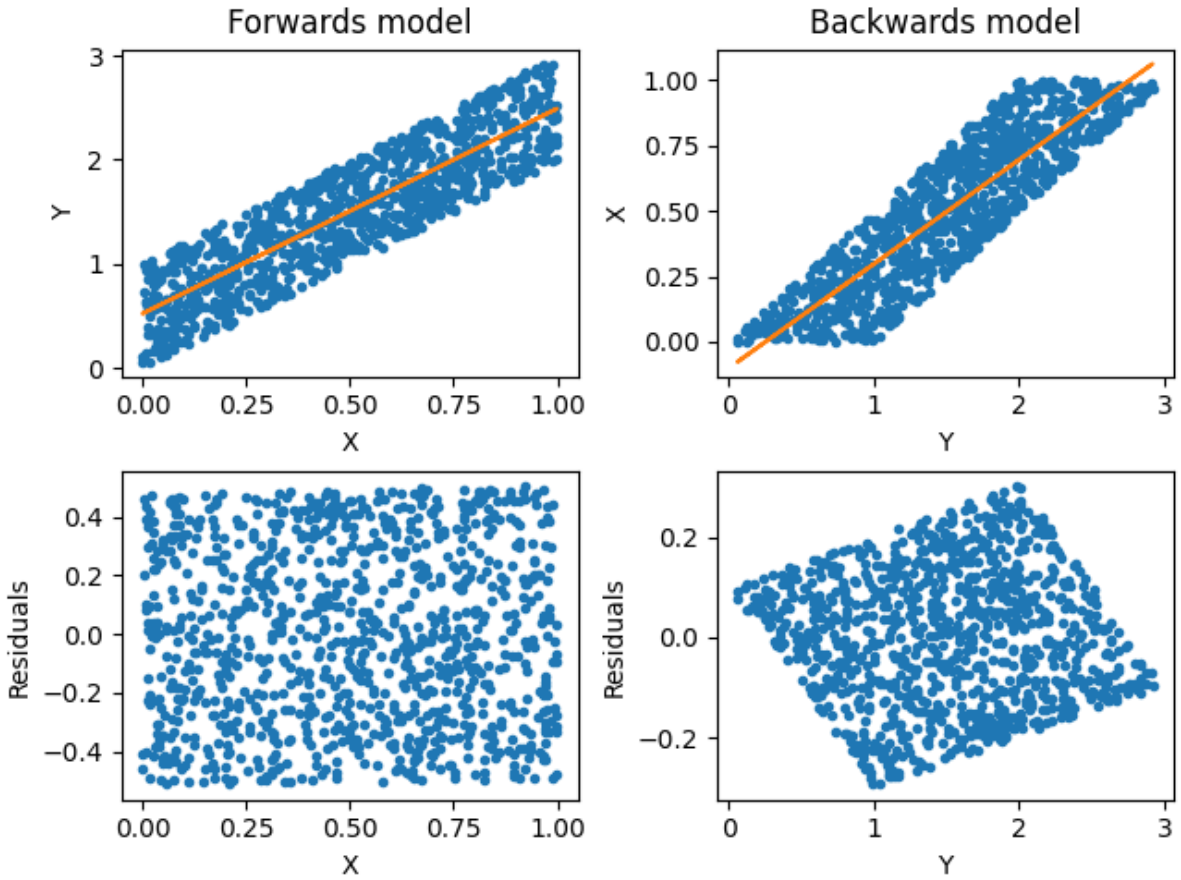
if and only if ξ_y and X are Gaussian.

It can be extended to multiple variables (see Shimizu et al. (2006), Peters et al. (2017)).

The linear non gaussian case

Example:

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



The linear non gaussian case

Ex from Hoyer et al. (2008)

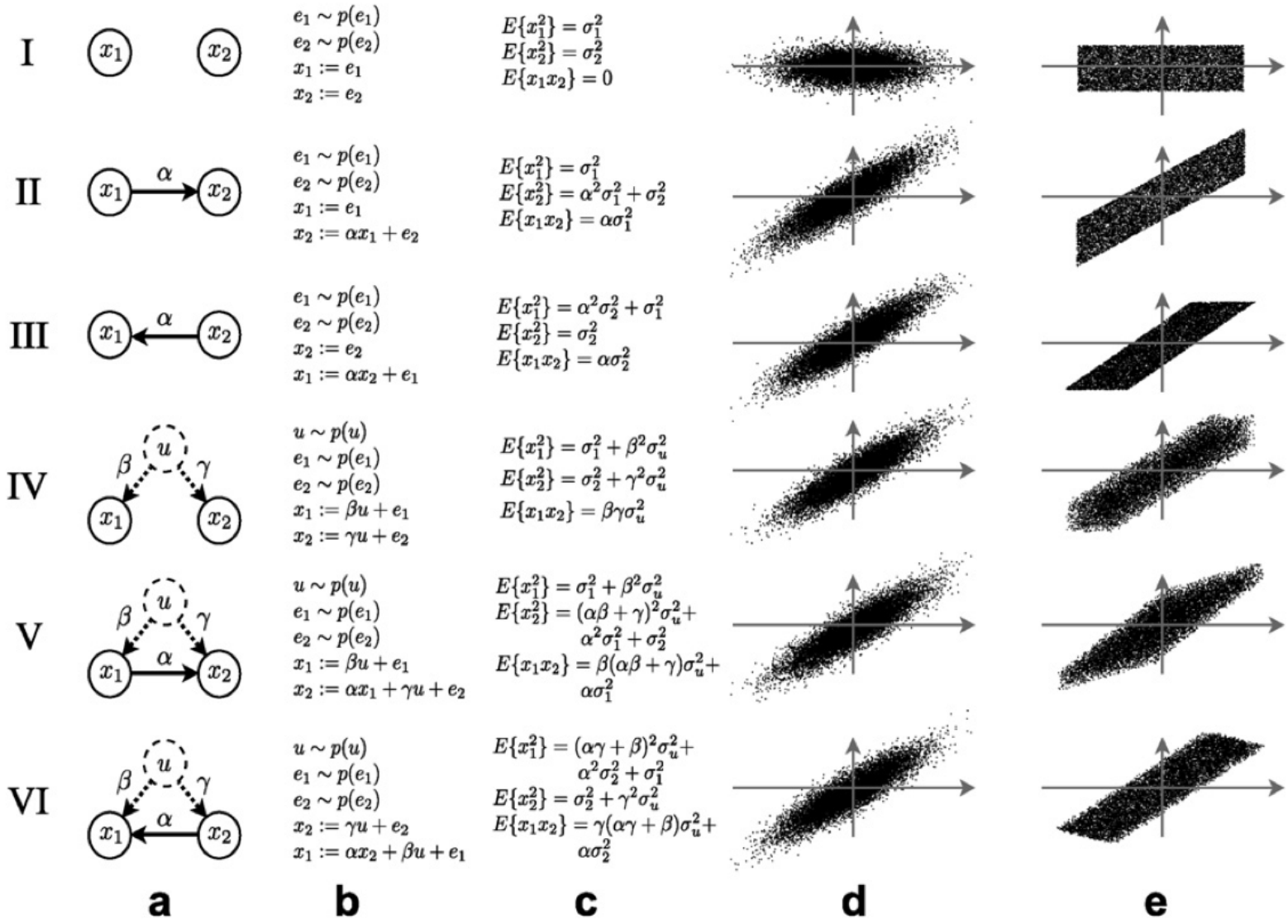


Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

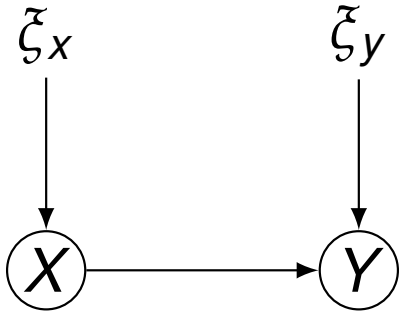
In practice

Multivariate case

Conclusion

The non linear ANM case

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 ANM case

Theorem (Identifiability of additive noise models, Hoyer et al, 2008)

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 f_X and random variable ξ_x such that

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

if and only if Complicated Condition is satisfied.

The non linear ANM case

Theorem (Identifiability of additive noise models, Hoyer et al, 2008)

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 f_X and random variable ξ_x such that

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

if and only if Complicated Condition is satisfied.

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

- The space that satisfy the condition is a 3-dimensional 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.

Be careful! the ANM is not closed under marginalization: with latent variables, if you assume ANM over $X \cup L$, then the marginal model over X may no longer be in the ANM class.

The non linear case

- The space that satisfy the condition is a 3-dimensional 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.

Be careful! the ANM is not closed under marginalization: with latent variables, if you assume ANM over $X \cup L$, then the marginal model over X may no longer be in the ANM class.

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

The post non linear case

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

Zhang and Hyvarinen (2009) have provided conditions under which this causal model is identifiable.

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

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 $\hat{f}_X$:

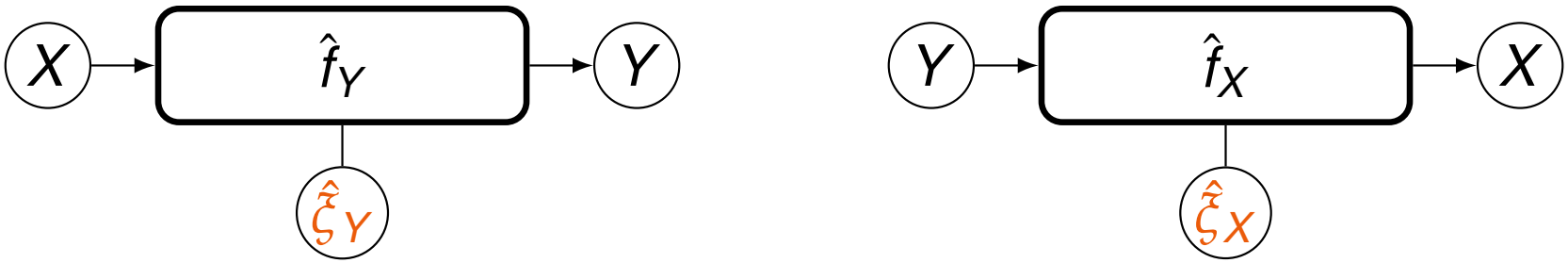
Causal order discovery procedure in the bivariate case

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 $\hat{\xi}_X$:

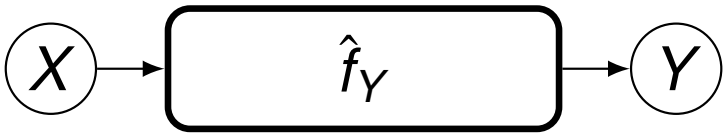


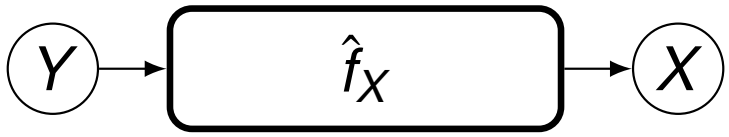
Causal order discovery procedure in the bivariate case

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

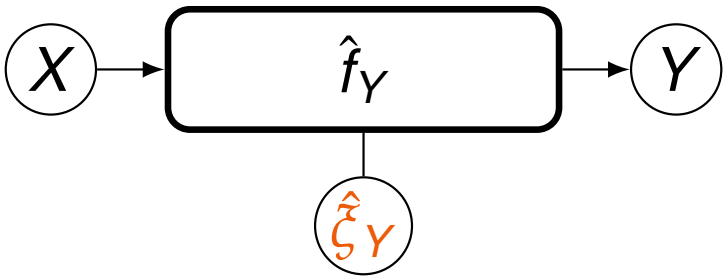
Procedure:

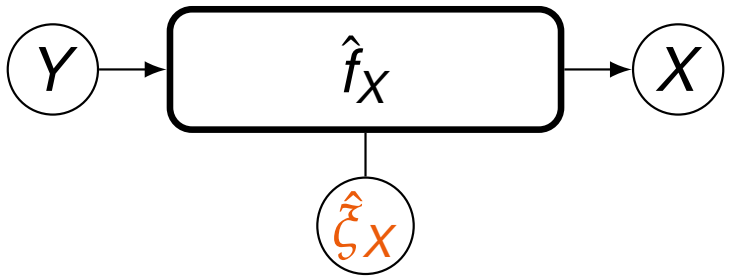
1. Fit $\hat{f}_Y$ and $\hat{f}_X$:





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 $\hat{I}(Y, \hat{\xi}_X) < \hat{I}(X, \hat{\xi}_Y)$

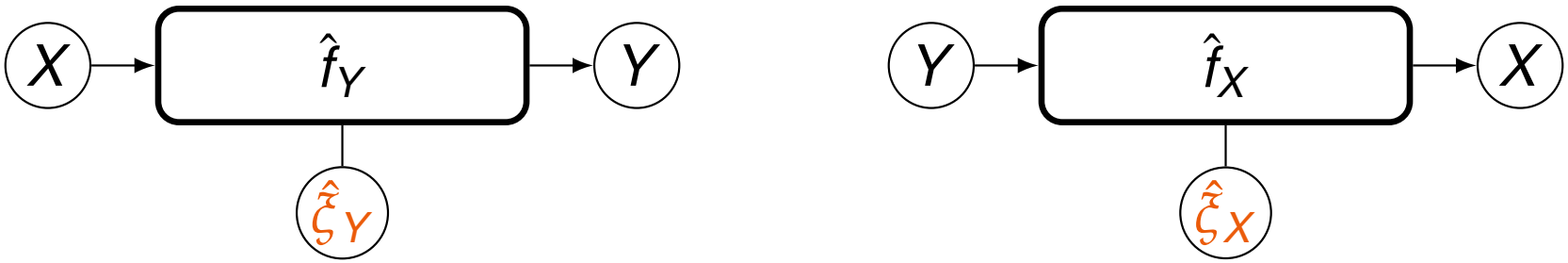
Causal order discovery procedure in the bivariate case

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

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

Assumptions

Causal sufficiency

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

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 $i < j$.

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 $\mathcal{G}$.

The graph does not contain dependencies not present in the observational data.

Assumptions

Causal sufficiency

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

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 $i < j$.

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 $\mathcal{G}$.
The graph does not contain dependencies not present in the observational data.

Assumptions

Causal sufficiency

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

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 $i < j$.

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 $\mathcal{G}$.
The graph does not contain dependencies not present in the observational data.

Remark: under Markov condition, faithfulness $\Longrightarrow$ minimality.

The algorithms

1. Topological ordering
2. Pruning

The algorithms

Algorithm 1 LiNGAM

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 $i \in S$ **do**
    for $j \in S \backslash \{i\}$ **do**
      $\hat{\xi}_{ij} = X_j - \frac{cov(X_i, X_j)}{var(X_i)} X_i$
    $h = \sum_{j \in S \backslash \{i\}} \hat{I}(X_i, \hat{\xi}_{ij})$
    $H = [H, h]$
  $i^* = arg\min_{i \in S} H$
  $S = S \backslash \{i^*\}$
  $\mathcal{T} = [\mathcal{T}, i^*]$
  $\forall j \in S, X_j = \hat{\xi}_{i^* j}$
until $|S| = 0$
Append($\mathcal{T}$, S_0)
Construct a strictly lower triangular matrix by following the order in $\mathcal{T}$, and estimate the connection strengths $a_{i,j}$ by using some conventional covariance-based regression.
if $a_{i,j} > 0$ **then**
  Add $X_i \to X_j$ to $\mathcal{G}$
Return $\mathcal{G}$

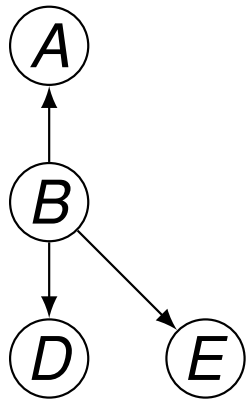
Algorithm 2 ANM

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 $\{X_i\}_{i \in S \backslash \{j\}}$
    $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$
    $h = \hat{I}(\{X_i\}_{i \in S \backslash \{j\}}, \hat{\xi}_{.j})$
    $H = [H, h]$
  $i^* = arg\min_{i \in S} H$
  $S = S \backslash \{i^*\}$
  $\mathcal{T} = [i^*, \mathcal{T}]$
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 $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \backslash \{i\}}$
    $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$
    if $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \backslash \{i\}} \not\perp\!\!\!\perp_P \xi_{.j}$ **then**
      Add $X_i \to X_j$ to $\mathcal{G}$
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 $D, E \mapsto B$ and $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$
- Estimate $B, E \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(\{B, E\}, \hat{\xi}_d)$

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

ANM in action (3/4)

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

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

- 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 $\hat{\xi}_b$
  - $H_3 = \hat{I}(\{D, E\}, \hat{\xi}_b)$

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

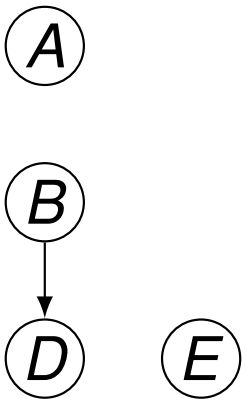
- Estimate $D \mapsto B$ and $\hat{\xi}_b$
  - $H_1 = \hat{I}(D, \hat{\xi}_b)$
- Estimate $B \mapsto D$ and $\hat{\xi}_d$
  - $H_2 = \hat{I}(B, \hat{\xi}_d)$

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

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

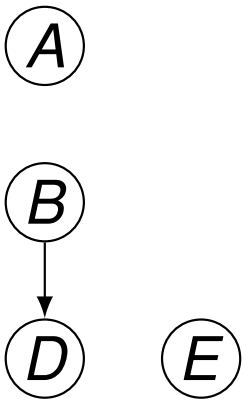
ANM in action (4/4)

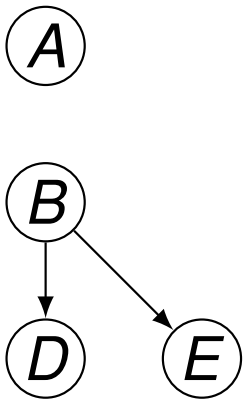
$$\mathcal{T} = [B, D, E, A]$$



ANM in action (4/4)

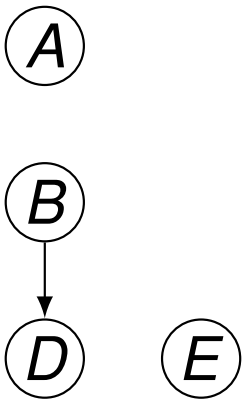
$$\mathcal{T} = [B, D, E, A]$$

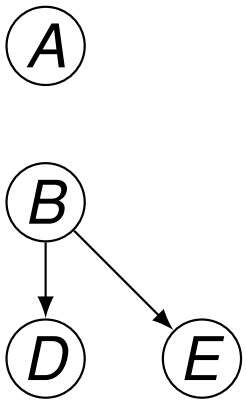


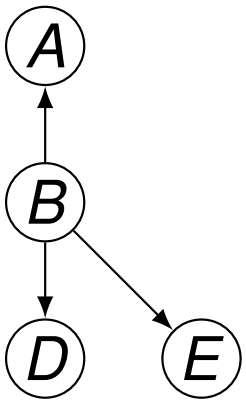


ANM in action (4/4)

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







Exercise 1

- 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?
  And what about PC?

Exercise 1

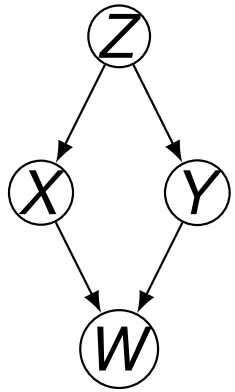
- Suppose the true graph on right;
- Assumptions: CMC, causal sufficiency, minimality;
- Generative process:

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

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

Exercise 1

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

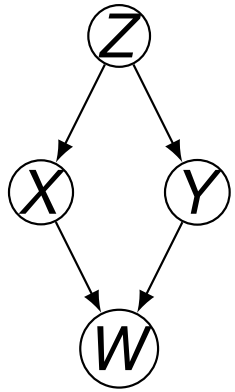


$$
\begin{aligned}
Z &= \xi_z & \xi_z &\sim U(0,1);\\
X &= a * Z + \xi_x & \xi_x &\sim U(0,1);\\
Y &= b * Z + \xi_y & \xi_y &\sim U(0,1);\\
W &= c * X - d * Y + \xi_w & \xi_w &\sim N(0,1).
\end{aligned}
$$

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

Exercise 2

- 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?

Exercise 3

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

Table of content

Reminder

Problem statement

The linear case

The non linear ANM case

The post non linear case

In practice

Multivariate case

Conclusion

Conclusion

- Noise-based methods can discover the causal graph
  - Under linear non gaussian models
  - Under non-linear additive noise models
- Advantages:
  - Can discovery the true graph;
  - Faithfulness is not needed.
- Drawbacks:
  - Semi parametric assumptions;
  - Need large sample size.
- Extensions
  - Without causal sufficiency if linear relations;
  - With cyclic graphs;
  - Extension to discrete additive noise models;
  - Time series.

Conclusion

- Noise-based methods can discover the causal graph
  - Under linear non gaussian models
  - Under non-linear additive noise models
- Advantages:
  - Can discovery the true graph;
  - Faithfulness is not needed.
- Drawbacks:
  - Semi parametric assumptions;
  - Need large sample size.
- Extensions
  - Without causal sufficiency if linear relations;
  - With cyclic graphs;
  - Extension to discrete additive noise models;
  - Time series.

Conclusion

- Noise-based methods can discover the causal graph
  - Under linear non gaussian models
  - Under non-linear additive noise models
- Advantages:
  - Can discovery the true graph;
  - Faithfulness is not needed.
- Drawbacks:
  - Semi parametric assumptions;
  - Need large sample size.
- Extensions
  - Without causal sufficiency if linear relations;
  - With cyclic graphs;
  - Extension to discrete additive noise models;
  - Time series.

References (1/2)

Direct inspirations

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. Shimizu, 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 (2/2)

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