CSDS 452 Causality and Machine Learning

Lecture 8: Causal Discovery

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Recap: Identification and Estimation

- Two components in learning causality
  - (1) Identification
  - (2) Estimation, inference

Underlying Causality

Identifying assumptions   Identification

Population distribution

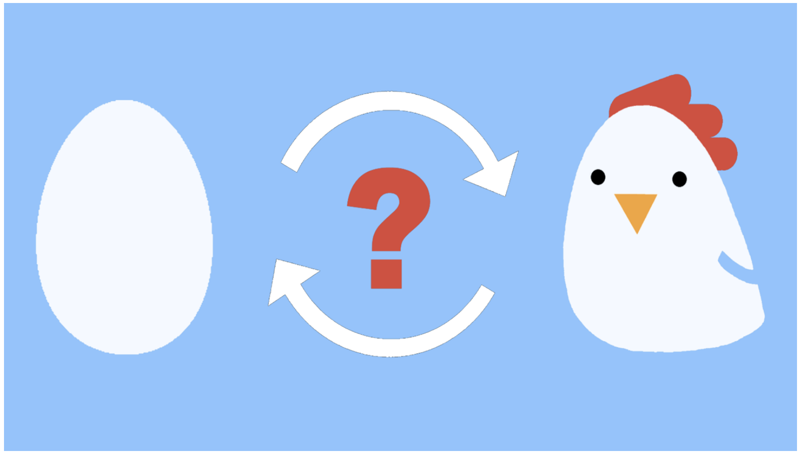
Sampling   Estimation, inference

Observational data

Recap: Classical causal effect estimation methods

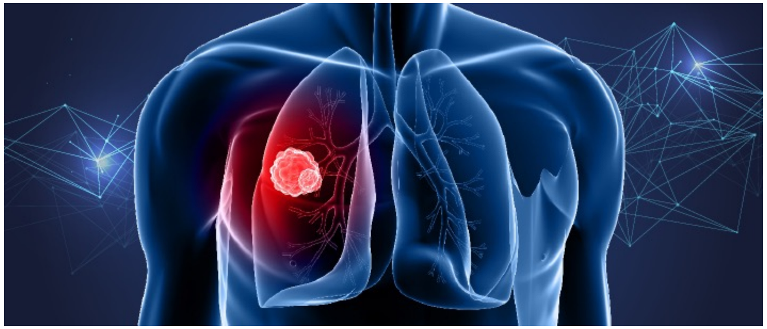
- Re-weighting methods
- Stratification methods
- Matching methods
- Tree-based methods
- Regression Adjustment
- Subspace Learning
- Difference-in-Difference
- ……

Two Main Problems in Causal Inference



Causal discovery

What causally affects what?



Causal effect estimation How significant is the causal effect?

How can we usually get a causal graph?

Common knowledge

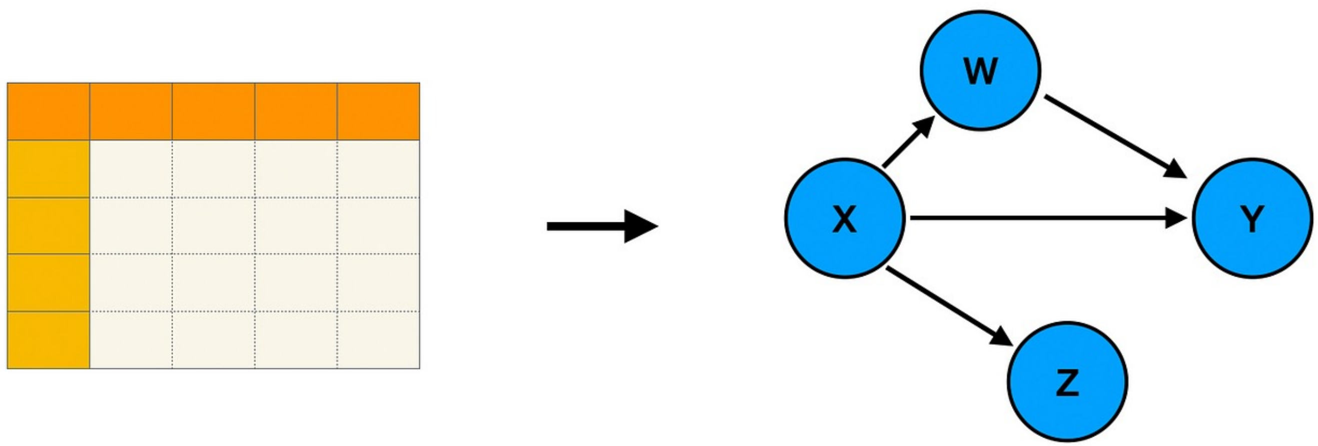
Domain experts

What if we don't have a given causal graph?

Causal Discovery

Data

Causal model



https://towardsdatascience.com/causal-discovery-6858f9af6dcb

Causal discovery aims to **find causal relations from data**. In other words, given a dataset, *derive* a causal model that describes it.

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Faithfulness Assumption

Recall the Markov assumption:

$$X \perp\!\!\!\perp_G Y \mid Z \quad \Rightarrow \quad X \perp\!\!\!\perp_P Y \mid Z$$

Causal graph $\longrightarrow$ Data

Causal graph $\xleftarrow{?}$ Data

Faithfulness Assumption

Recall the Markov assumption:

$$X \perp\!\!\!\perp_G Y \mid Z \quad \Rightarrow \quad X \perp\!\!\!\perp_P Y \mid Z$$

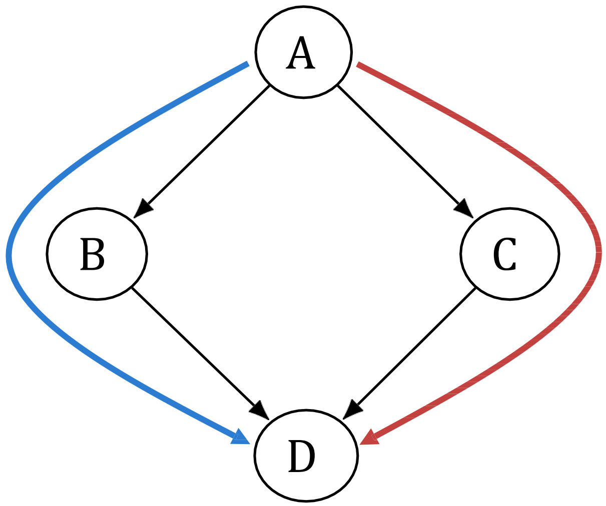
$$\text{Causal graph} \longrightarrow \text{Data}$$

$$\text{Causal graph} \longleftarrow \text{Data}$$

Faithfulness: $X \perp\!\!\!\perp_G Y \mid Z \quad \Leftarrow \quad X \perp\!\!\!\perp_P Y \mid Z$

Example: Violation of Faithfulness

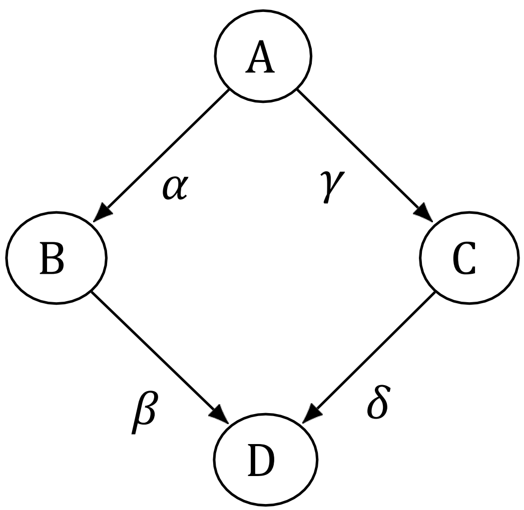
Faithfulness: $X \perp\!\!\!\perp_P Y \mid Z \;\Rightarrow\; X \perp\!\!\!\perp_G Y \mid Z$



A $\perp\!\!\!\perp$ D, but A and D are not d-separated

Example: Violation of Faithfulness

Faithfulness: $X \perp\!\!\!\perp_P Y | Z \;\Rightarrow\; X \perp\!\!\!\perp_G Y | Z$



$$B := \alpha A$$
$$C := \gamma A$$
$$D := \beta B + \delta C$$

$$D := (\alpha\beta + \gamma\delta)A$$

If $\alpha\beta + \gamma\delta = 0$, we have A $\perp\!\!\!\perp$ D in population

A $\perp\!\!\!\perp$ D, but A and D are not d-separated

Causal Sufficiency and Acyclicity

- **Causal Sufficiency**: there are no unobserved confounders of any of the variables in the graph.

- **Acyclicity**: there are no cycles in the graph.

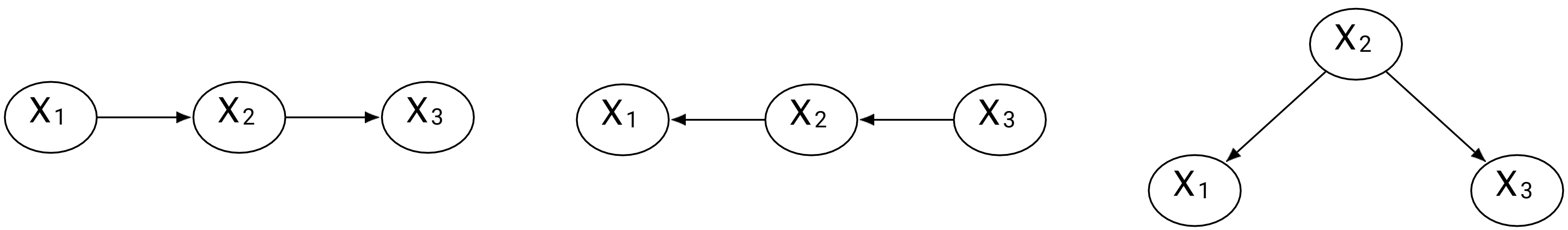
Important assumptions for now

- Markov assumption
- Faithfulness assumption
- Causal sufficiency assumption
- Acyclicity assumption

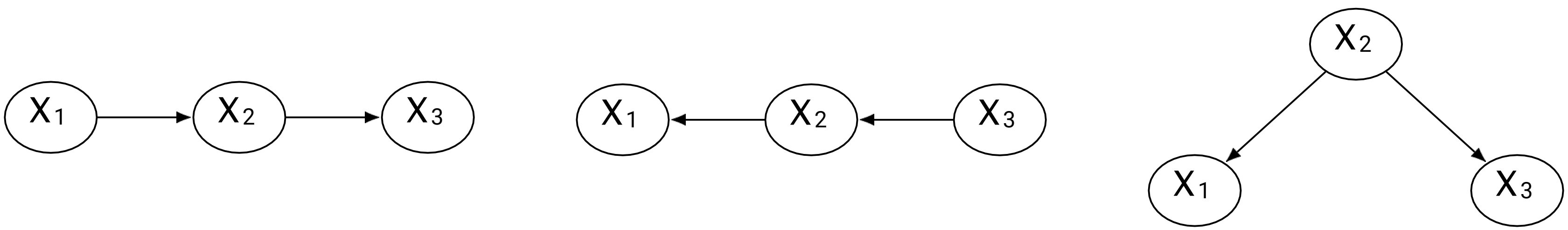
Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Chains and Forks Encode Same Independencies



Chains and Forks Encode Same Independencies



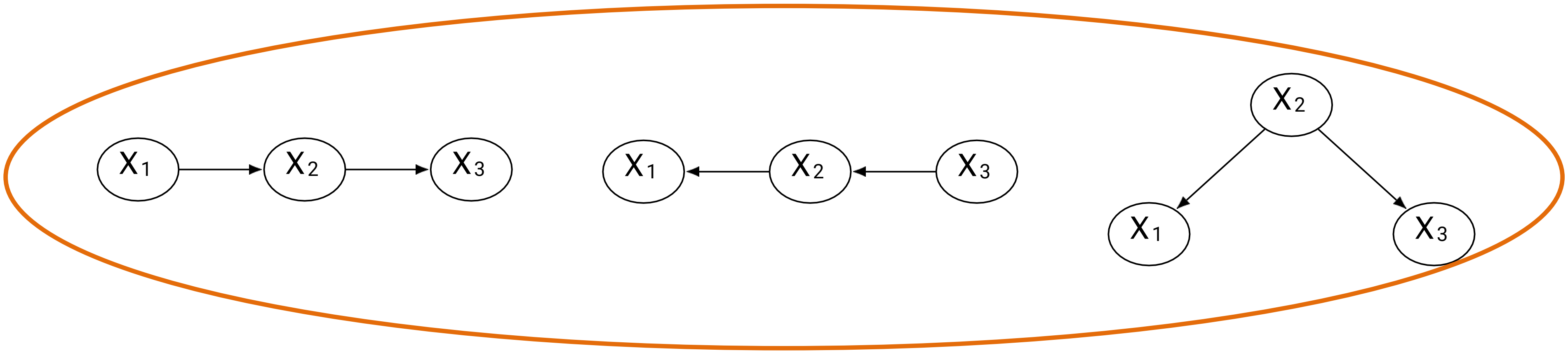
$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

$X_1 \not\perp\!\!\!\perp X_2$ and $X_2 \not\perp\!\!\!\perp X_3$

$$X_1 \not\perp\!\!\!\perp X_3$$

Chains and Forks Encode Same Independencies

Markov equivalent (they are in the same **Markov equivalence class**)



$X_1 \perp\!\!\!\perp X_3 | X_2$

$X_1 \not\perp\!\!\!\perp X_2$ and $X_2 \not\perp\!\!\!\perp X_3$

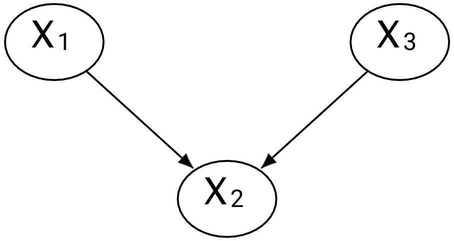
$X_1 \not\perp\!\!\!\perp X_3$

Markov equivalence class: a set that contains all Directed Acyclic Graphs (DAGs) that share the same conditional independence structure

Special cases

- In some cases, we can get many causal graphs which are in the same Markov equivalence class.
- But some cases are special!

Immoralities on Graph



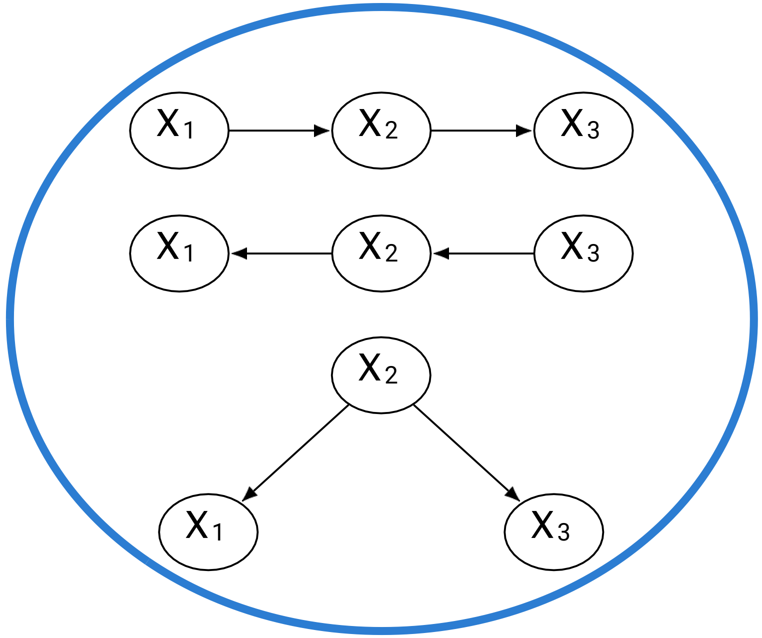
Blocked path

A graph immorality occurs when a child has two or more parents that without connection between them

Immoralities are Special

$X_1 \perp\!\!\!\perp X_3 | X_2$

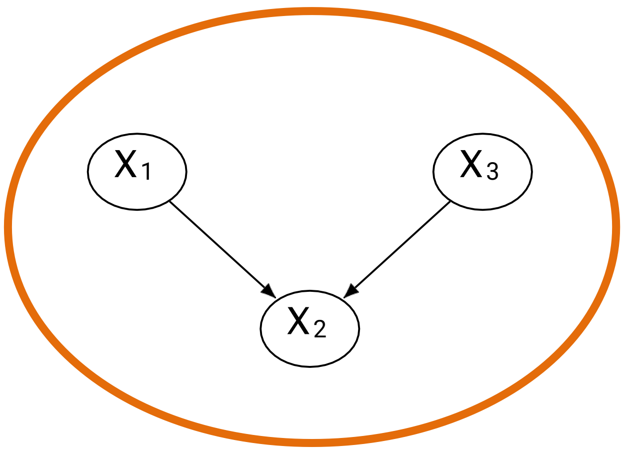
$X_1 \not\perp\!\!\!\perp X_3$



Immoralities are Special

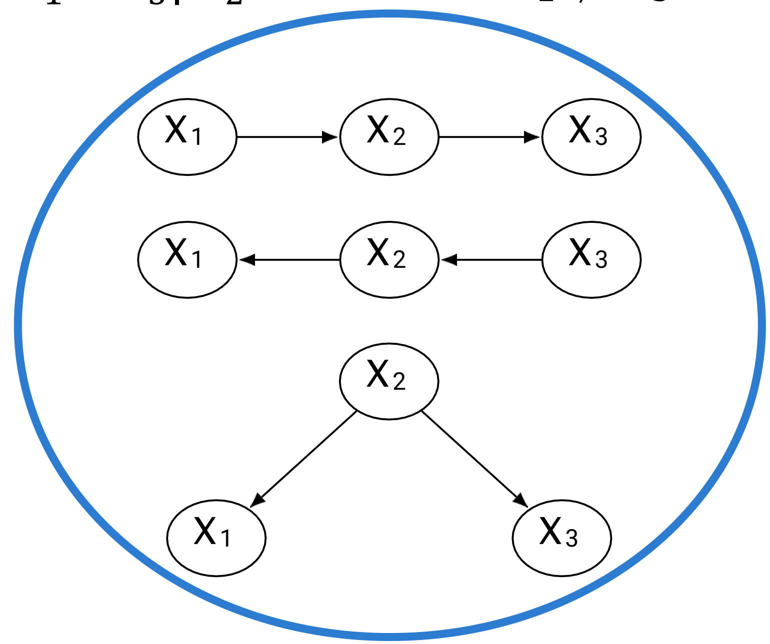
Markov equivalence class 1

$X_1 \perp\!\!\!\perp X_3$ $\quad$ $X_1 \not\perp\!\!\!\perp X_3 | X_2$



Markov equivalence class 2

$X_1 \perp\!\!\!\perp X_3 | X_2$ $\quad$ $X_1 \not\perp\!\!\!\perp X_3$



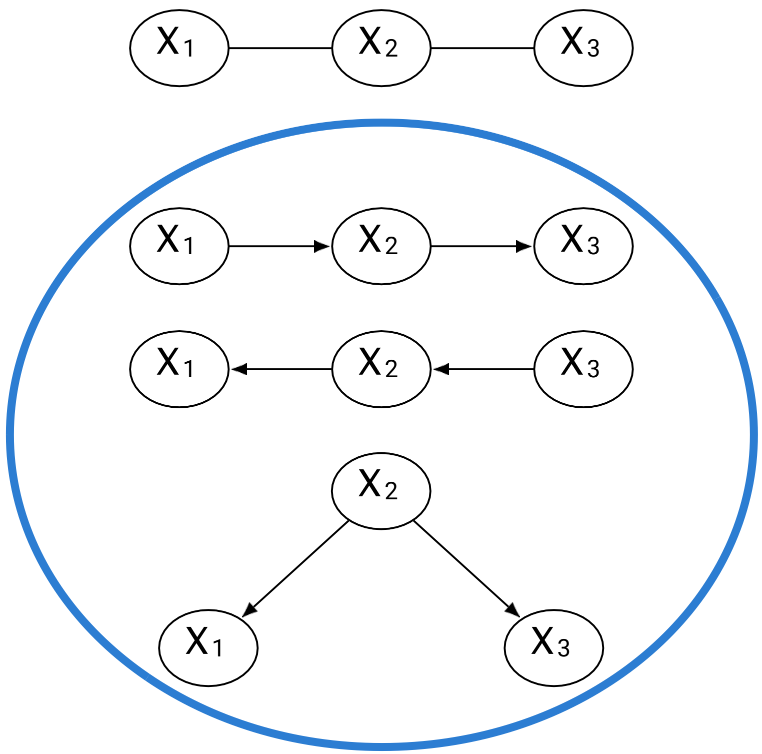
- Immorality is alone in its own Markov equivalence class!
- Therefore, compared with chains and forks, it is easier to identify immorality based on conditional independence among variables.

What can we discover from chains and forks?

Skeletons

- Skeleton: transform the directed edges into undirected ones
- The skeleton tell us that

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

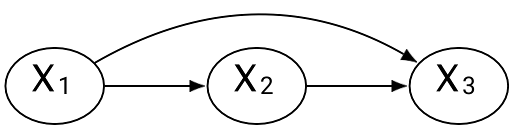


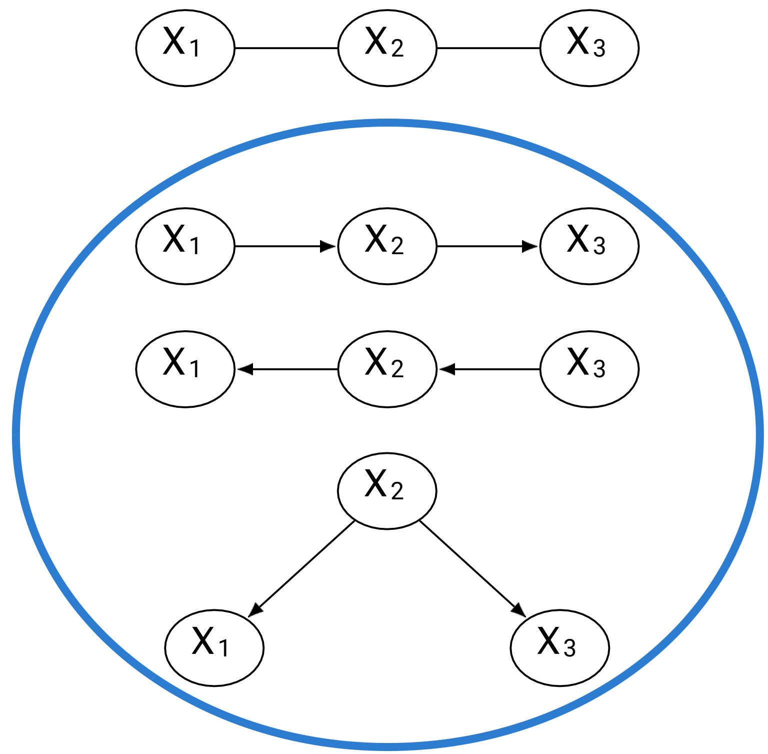
Skeletons

- Skeleton: transform the directed edges into undirected ones
- The skeleton tell us that

$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

$$X_1 \not\perp\!\!\!\perp X_3 \mid X_2$$

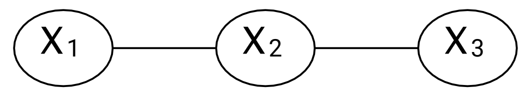


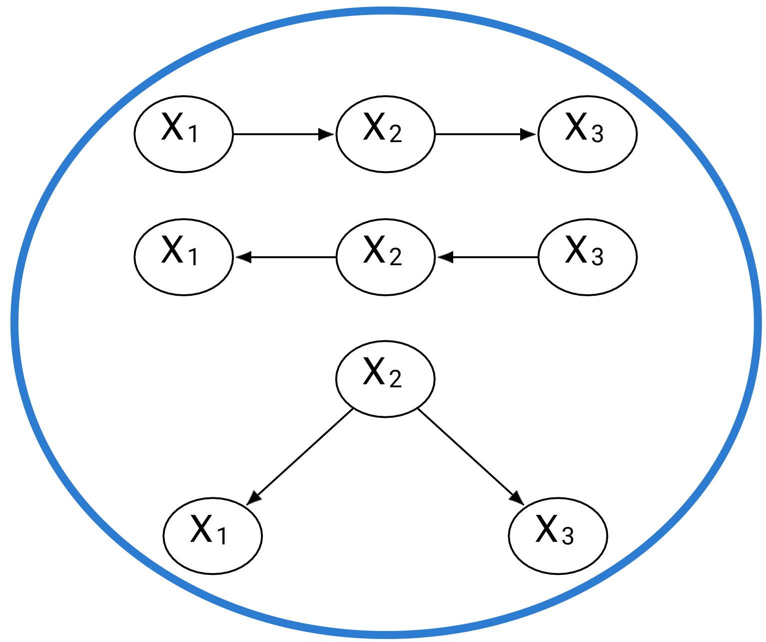


Skeletons

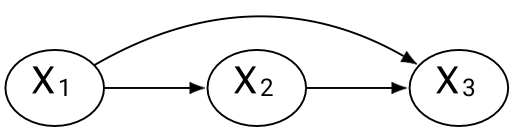
- Skeleton: transform the directed edges into undirected ones
- The skeleton tell us that

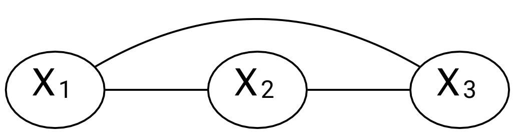
$$X_1 \perp\!\!\!\perp X_3 | X_2$$





$$X_1 \not\perp\!\!\!\perp X_3 | X_2$$





Markov Equivalence via Immoral Skeletons

Two important things that we can use to distinguish graphs:

1. Immoralities
2. Skeleton

Theorem: Two graphs are Markov equivalent if and only if they have the same skeleton and same immoralities.[1]

[1] Verma T S, Pearl J. On the equivalence of causal models[J]. arXiv preprint arXiv:1304.1108, 2013.

Markov Equivalence via Immoral Skeletons

- Based on conditional independencies, the best we can do is to discovery “**essential graph**”(a.k.a., completely partial DAG, **CPDAG**), i.e., skeleton + immoralities

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

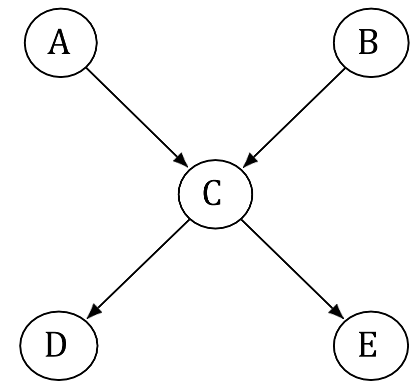
The PC Algorithm

- Start with complete undirected graph
- 3 steps

The PC Algorithm

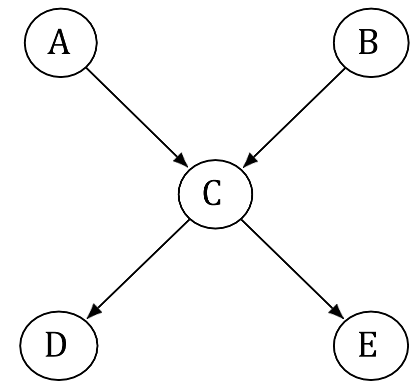
- Start with complete undirected graph
- 3 steps

True causal graph



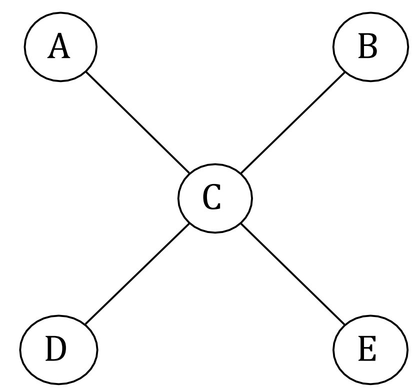
The PC Algorithm

True causal graph



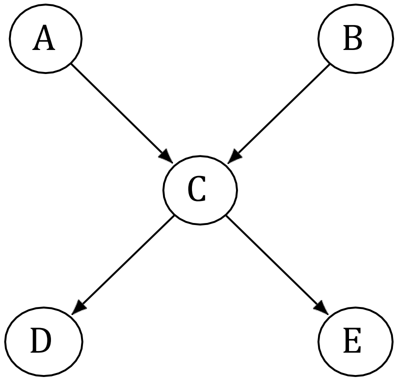
- Start with complete undirected graph
- 3 steps

1. Identify the skeleton



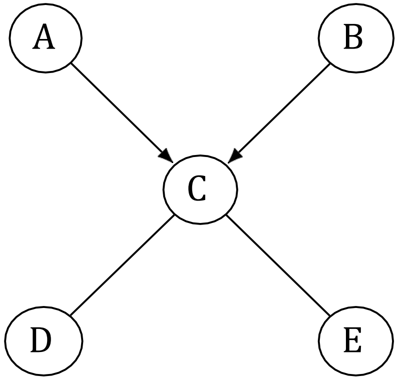
The PC Algorithm

True causal graph



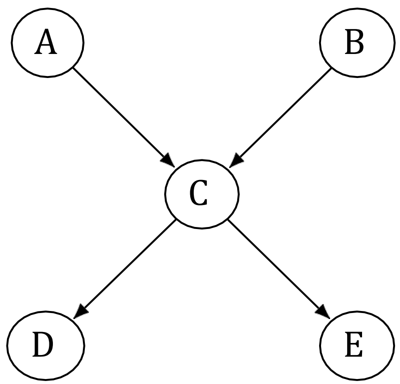
- Start with complete undirected graph
- 3 steps

1. Identify the skeleton
2. Identify immoralities and orient them



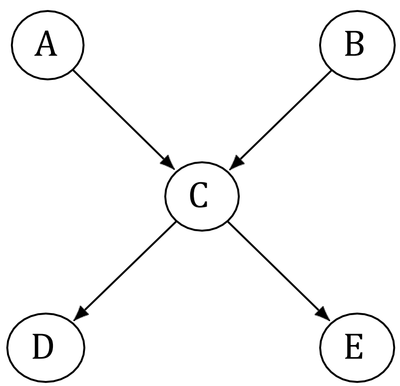
The PC Algorithm

True causal graph



- Start with complete undirected graph
- 3 steps

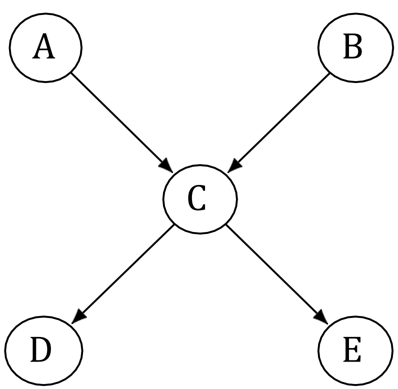
1. Identify the skeleton
2. Identify immoralities and orient them
3. Orient more edges incident on colliders (based on the fact that we would have identified more immoralities in Step 2)



Step 1: Identifying the Skeleton

- Start with complete undirected graph
- Remove edges $X - Y$ where $X \perp\!\!\!\perp Y \mid Z$ for some (potentially empty) conditioning set Z, starting with the empty conditioning set and increasing the size

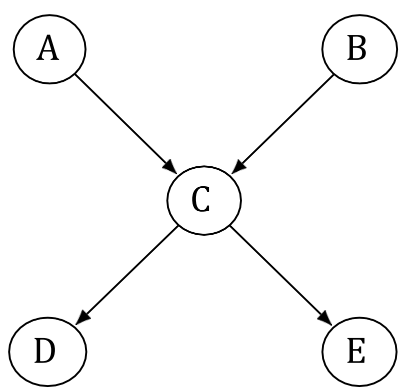
True causal graph



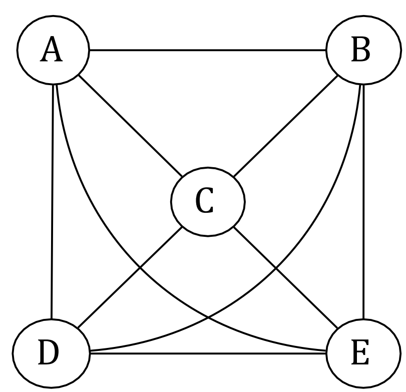
Step 1: Identifying the Skeleton

- Start with complete undirected graph
- Remove edges $X - Y$ where $X \perp\!\!\!\perp Y \mid Z$ for some (potentially empty) conditioning set Z, starting with the empty conditioning set and increasing the size

True causal graph



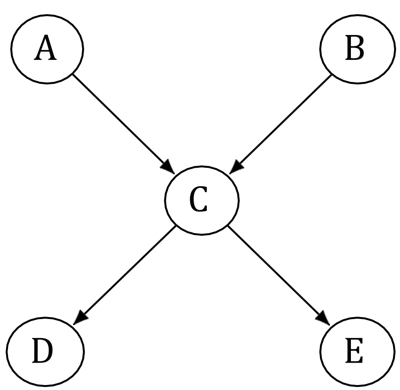
complete undirected graph

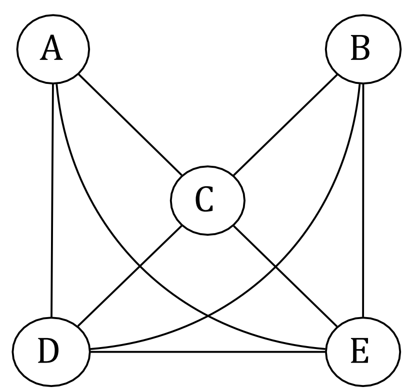


Step 1: Identifying the Skeleton

- Start with complete undirected graph
- Remove edges $X - Y$ where $X \perp\!\!\!\perp Y |\, Z$ for some (potentially empty) conditioning set Z, starting with the empty conditioning set and increasing the size

True causal graph



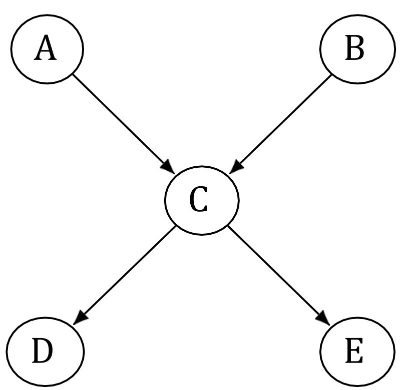


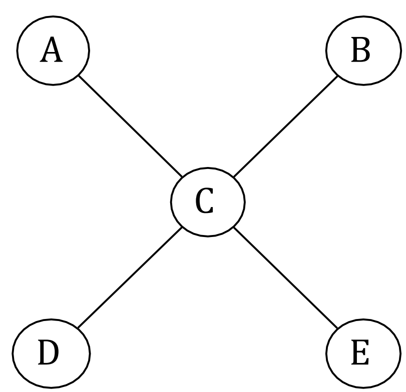
$\mathrm{A} \perp\!\!\!\perp \mathrm{B} |\, \{\}$

Step 1: Identifying the Skeleton

- Start with complete undirected graph
- Remove edges $X - Y$ where $X \perp\!\!\!\perp Y \mid Z$ for some (potentially empty) conditioning set Z, starting with the empty conditioning set and increasing the size

True causal graph





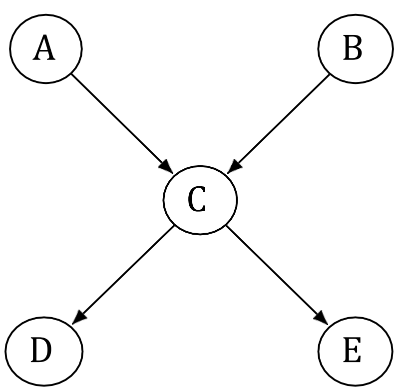
$$A \perp\!\!\!\perp B \mid \{\}$$

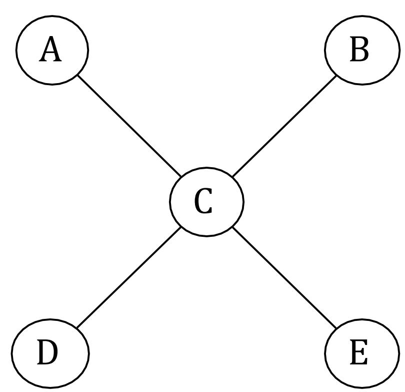
$$\forall (X, Y), X \perp\!\!\!\perp Y \mid \{C\}$$

Step 2: Identifying the Immoralities

- Now for any paths X – Z – Y in the current graph where the following are true:
    1. In Step 1, we discovered that there is no edge between X and Y.
    2. Z was not in the conditioning set that makes X and Y conditionally independent.
- Then, we know X – Z – Y forms an immortality

True causal graph

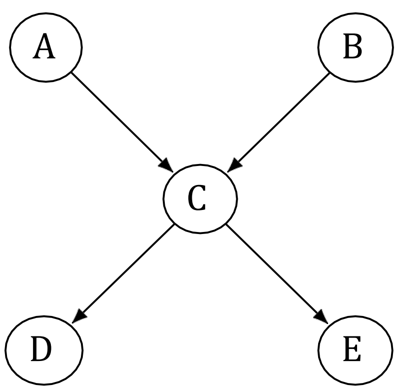


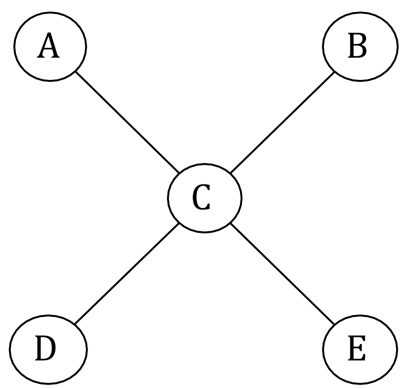


Step 2: Identifying the Immoralities

- Now for any paths X – Z – Y in the current graph where the following are true:
  1. In Step 1, we discovered that there is no edge between X and Y.
  2. Z was not in the conditioning set that makes X and Y conditionally independent.
- Then, we know X – Z – Y forms an immortality

True causal graph





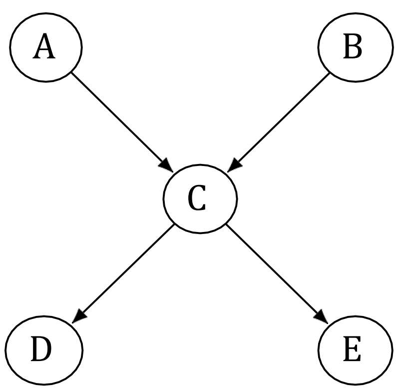
1. A – B has no edge between them.
2. C was not in the conditioning set that makes A and B conditionally independent.

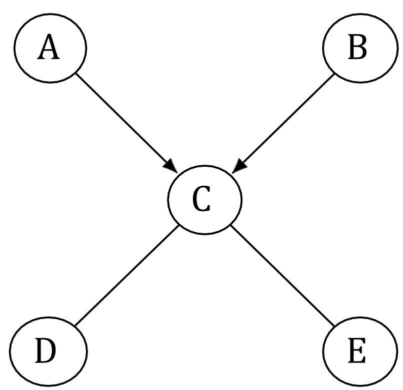
$$A \perp\!\!\!\perp B \mid \{\}$$

Step 2: Identifying the Immoralities

- Now for any paths X – Z – Y in the current graph where the following are true:
  1. In Step 1, we discovered that there is no edge between X and Y.
  2. Z was not in the conditioning set that makes X and Y conditionally independent.
- Then, we know X – Z – Y forms an immortality

True causal graph



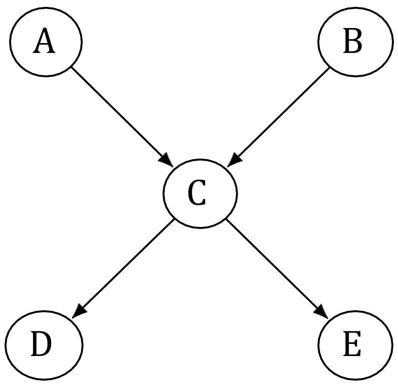


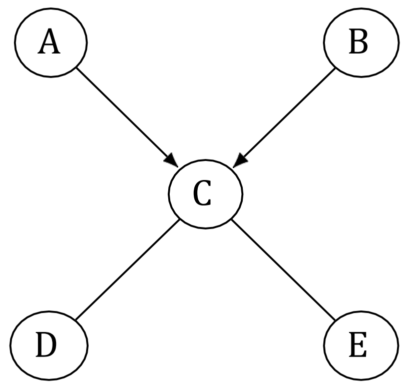
1. A – B has no edge between them.
2. C was not in the conditioning set that makes A and B conditionally independent.

$$A \perp\!\!\!\perp B \mid \{\}$$

Step 3: Orienting Qualifying Edges Incident on Colliders

True causal graph

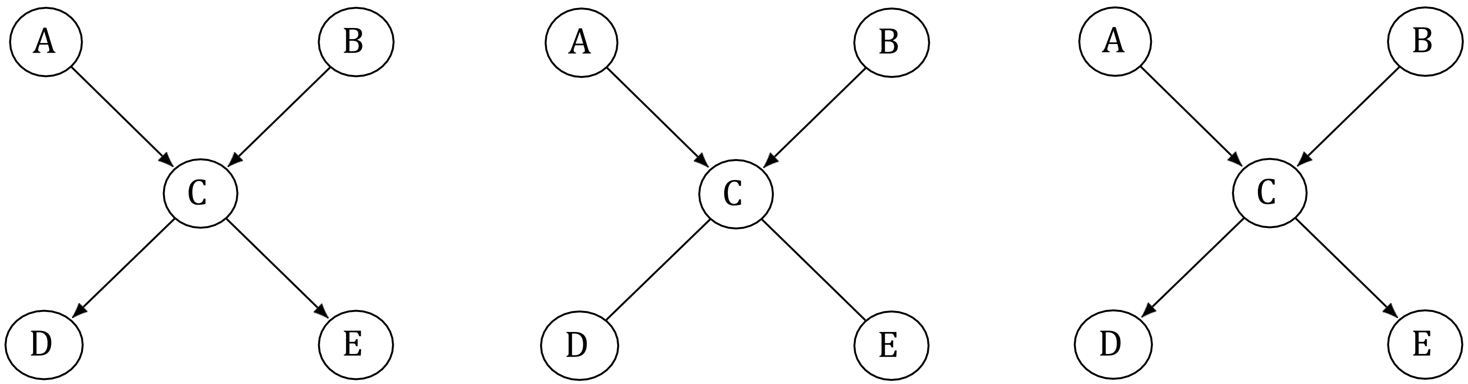




Key: orient more edges by using the fact that we should have discovered all immoralities in Step 2.

Step 3: Orienting Qualifying Edges Incident on Colliders

True causal graph



Key: orient more edges by using the fact that we should have discovered all immoralities in Step 2.

For any edge Z—Y which is part of a partially directed path X→Z—Y, where there is no edge connecting X and Y can be oriented as Z →Y
(otherwise, X→Z←Y is immoral and should be discovered in Step 2)

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Weaker Assumptions?

- No assumed causal sufficiency: Fast Casual Inference (FCI) algorithm (Spirtes et al., 2001 [1])
- No assumed acyclicity: CCD algorithm (Richardson, 1996 [2])
- Neither causal sufficiency nor acyclicity: SAT-based causal discovery (Hyttinen et al., 2013 [3])

[1] Spirtes P, Glymour C N, Scheines R. Causation, prediction, and search[M]. MIT press, 2000.

[2] Richardson T. Feedback models: Interpretation and discovery[D]. Ph. D. thesis, Carnegie Mellon, 1996.

[3] Hyttinen A, Hoyer P O, Eberhardt F, et al. Discovering cyclic causal models with latent variables: A general SAT-based procedure[J]. arXiv preprint arXiv:1309.6836, 2013.

Hardness of Conditional Independence Testing

- Independence-based causal discovery algorithms rely on accurate conditional independence testing.
- Conditional independence testing is simple if we have <u>infinite</u> data, however, it is much harder in finite data, and it can sometimes require a lot of data to get accurate test results.

Can We Do Better?

- With faithfulness, we saw we can identify the essential graph (Markov equivalence class).
- If we have multinomial distributions (Meek, 1995) or linear Gaussian structural equations (Geiger & Pearl, 1988), we can only identify a graph up to its Markov equivalence class.

Meek C. Strong completeness and faithfulness in Bayesian networks[J]. arXiv preprint arXiv:1302.4973, 2013.

Geiger D, Pearl J. On the logic of causal models[M]//Machine Intelligence and Pattern Recognition. North-Holland, 1990, 9: 3-14.

Can We Do Better?

- With faithfulness, we saw we can identify the essential graph (Markov equivalence class).

- If we have multinomial distributions (Meek, 1995) or linear Gaussian structural equations (Geiger & Pearl, 1988), we can only identify a graph up to its Markov equivalence class.

What about non-Gaussian structural equations? Or nonlinear structural equations?

Meek C. Strong completeness and faithfulness in Bayesian networks[J]. arXiv preprint arXiv:1302.4973, 2013.

Geiger D, Pearl J. On the logic of causal models[M]//Machine Intelligence and Pattern Recognition. North-Holland, 1990, 9: 3-14.

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Issues with Independence-Based Causal Discovery

- Requires faithfulness assumption
- Large samples can be necessary for conditional independence tests
- Only identifies the Markov equivalence class

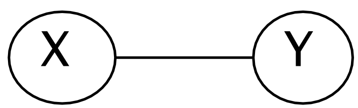
Two Variable Case: Markov Equivalence

- Infinite data: $P(x, y)$

Two Variable Case: Markov Equivalence

- Infinite data: $P(x, y)$

Essential graph:



Two Variable Case: SCMs Perspective

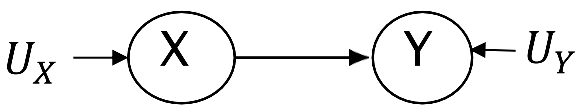
- Proposition: For every joint distribution $P(x, y)$ on two real-valued random variables, there is an SCM in either direction that generates data consistent with $P(x, y)$.

Two Variable Case: SCMs Perspective

- Proposition: 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 $P(x, y)$.

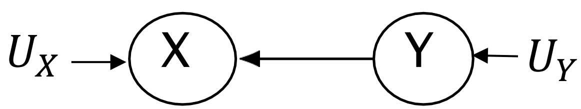
Mathematically, there exists a function f_Y such that

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



And there exists a function f_X such that

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



Without any assumption, we cannot distinguish these two causal graphs

We must make assumptions about the parametric form.

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Linear Non-Gaussian Assumption

- Previous study shows that we cannot hope to identify the graph more precisely than the Markov equivalence class in the linear Gaussian noise setting [1]

- What if the noise is non-Gaussian?

[1] Geiger D, Pearl J. On the logic of causal models[M]//Machine Intelligence and Pattern Recognition. North-Holland, 1990, 9: 3-14.

Linear Non-Gaussian Assumption

All structural equations (causal mechanisms that generate the data) are of the following form:

$$Y := f(X) + U$$

where f is a linear function, $X \perp\!\!\!\perp U$, and U follows non-Gaussian distribution.

Identifiability in Linear Non-Gaussian Setting

Theorem [1]

In the linear non-Gaussian setting, if the true SCM is

$$Y \coloneqq f(X) + U, \qquad X \perp\!\!\!\perp U$$

then there does not exist an SCM in the reverse direction that can generate data consistent with $P(x, y)$.

[1] Shimizu S, Hoyer P O, Hyvärinen A, et al. A linear non-Gaussian acyclic model for causal discovery[J]. Journal of Machine Learning Research, 2006, 7(10).

Linear Non-Gaussian Identifiability Extensions

- Multivariate: Shimizu et al., (2006) [1]
- Drop causal sufficiency assumption: Hoyer et al. (2008) [2]
- Drop acyclicity assumption: Lacerda et al. (2008) [3]

[1] Shimizu S, Hoyer P O, Hyvärinen A, et al. A linear non-Gaussian acyclic model for causal discovery[J]. Journal of Machine Learning Research, 2006, 7(10).
[2] Hoyer P O, Shimizu S, Kerminen A J, et al. Estimation of causal effects using linear non-Gaussian causal models with hidden variables[J]. International Journal of Approximate Reasoning, 2008, 49(2): 362-378.
[3] Lacerda G, Spirtes P L, Ramsey J, et al. Discovering cyclic causal models by independent components analysis[J]. arXiv preprint arXiv:1206.3273, 2012.

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Identifiability in Nonlinear Additive Noise Setting

Recall: We cannot hope to identify the graph more precisely than the Markov equivalence class in the linear Gaussian noise setting

What if the structural equations are nonlinear?

Identifiability in Nonlinear Additive Noise Setting

Nonlinear additive noise assumption:

$$\forall i, X_i \coloneqq f_i(pa_i) + U_i$$

where f_i is nonlinear

Identifiability in Nonlinear Additive Noise Setting

Nonlinear additive noise assumption:

$$\forall i, X_i \coloneqq f_i(pa_i) + U_i$$

where f_i is nonlinear

Theorem (Hoyer et al. 2008): Under the Markov assumption, causal sufficiency, acyclicity, the nonlinear additive noise assumption, and a technical condition from Hoyer et al. (2008), we can identify the causal graph.

Hoyer, Patrik, et al. "Nonlinear causal discovery with additive noise models." Advances in neural information processing systems 21 (2008).

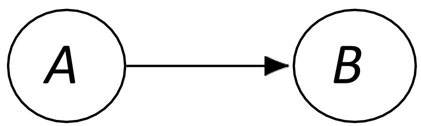
Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

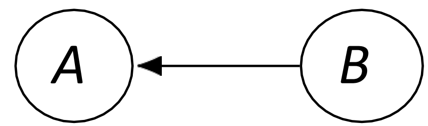
What if we can do interventions?

- Until now, our discussion of causal discovery is in the setting of observational data
- But what if we can do interventions (i.e., creating interventional data)?

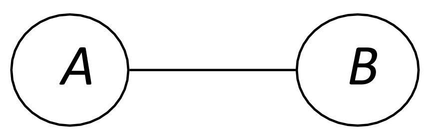
Consider 2-variable case



or



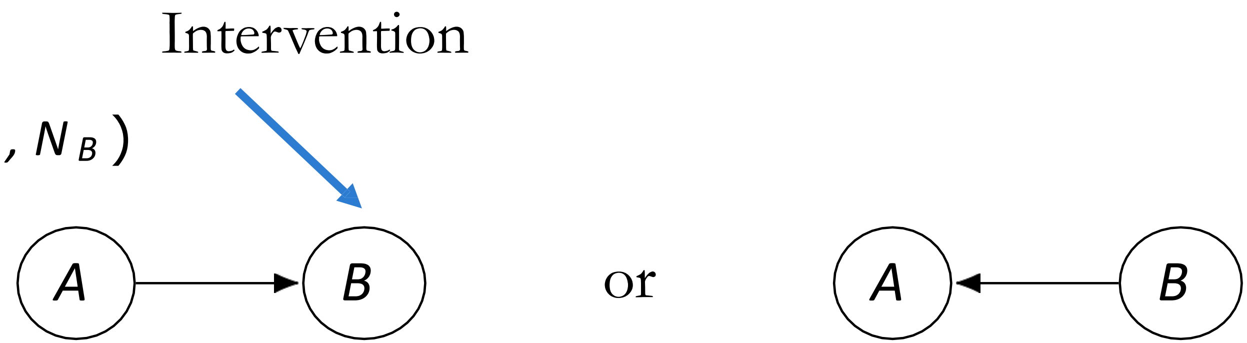
Same Markov equivalence class / essential graph



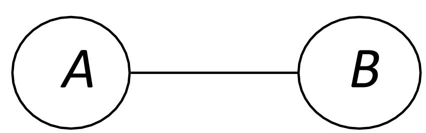
Intervention involved!

$A := N_A$

$B := f(A, N_B)$



Same Markov equivalence class / essential graph



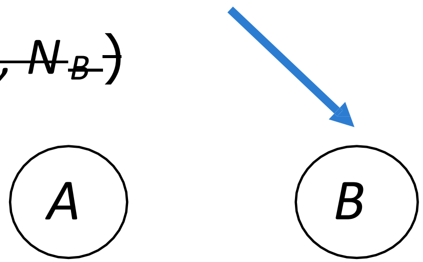
Intervention involved!

$A := N_A$

$\cancel{B := f(A, N_B)}$

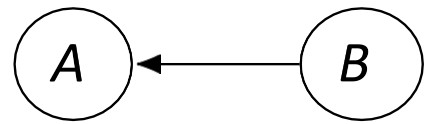
$B := N_B'$

Intervention

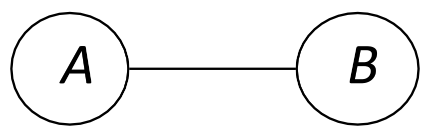


Manipulated graph

or



Same Markov equivalence class / essential graph



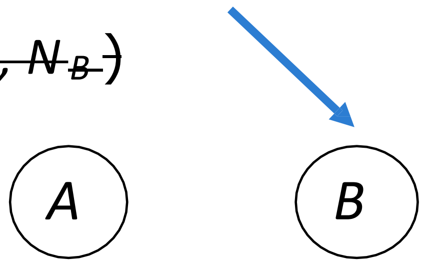
Intervention involved!

$A := N_A$

$\cancel{B := f(A, N_B)}$

$B := N_B^{'}$

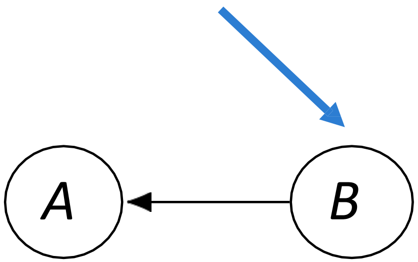
Intervention



Manipulated graph

or

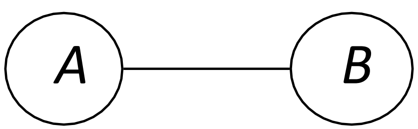
Intervention



$B := N_B$

$A := f(B, N_A)$

Same Markov equivalence class / essential graph



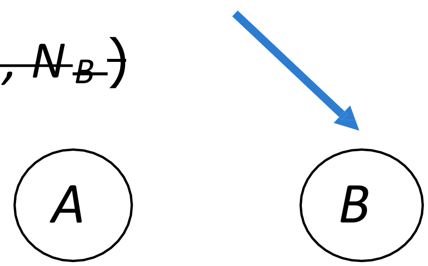
Intervention involved!

$A := N_A$

$\cancel{B := f(A, N_B)}$

$B := N_B'$

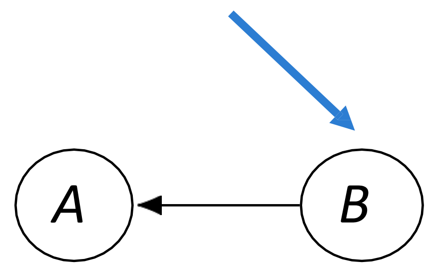
Intervention



Manipulated graph

or

Intervention

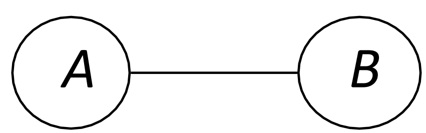


$B := N_B'$

$\cancel{B := N_B}$

$A := f(B, N_A)$

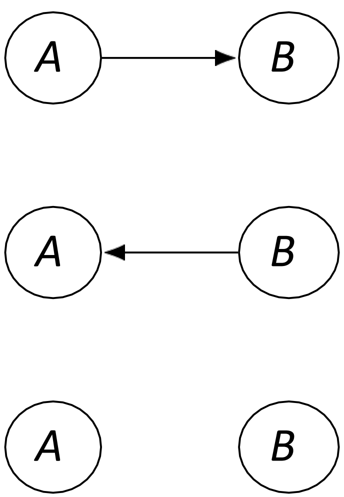
Same Markov equivalence class / essential graph



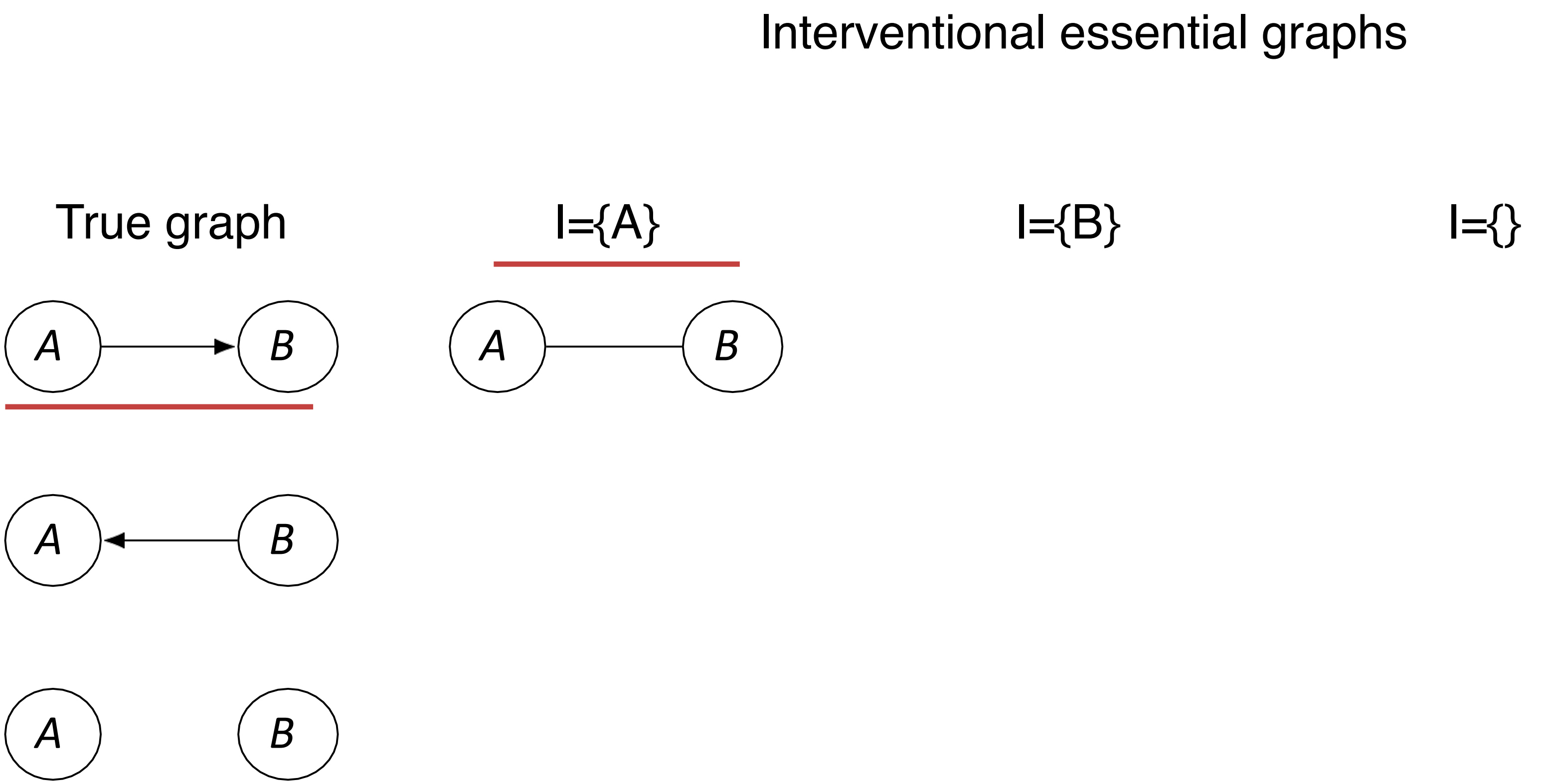
Two Variables: Interventional Essential Graphs

Interventional essential graphs

True graph | I={A} | I={B} | I={}



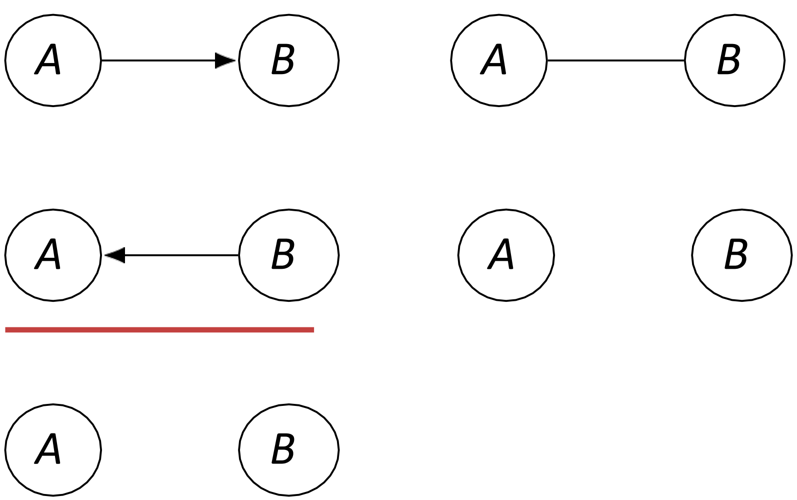
Two Variables: Interventional Essential Graphs



Two Variables: Interventional Essential Graphs

Interventional essential graphs

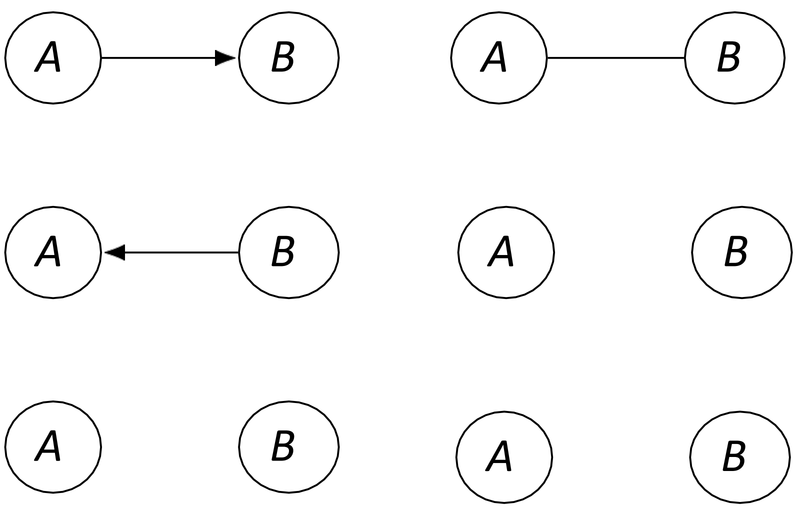
True graph | I={A} | I={B} | I={}



Two Variables: Interventional Essential Graphs

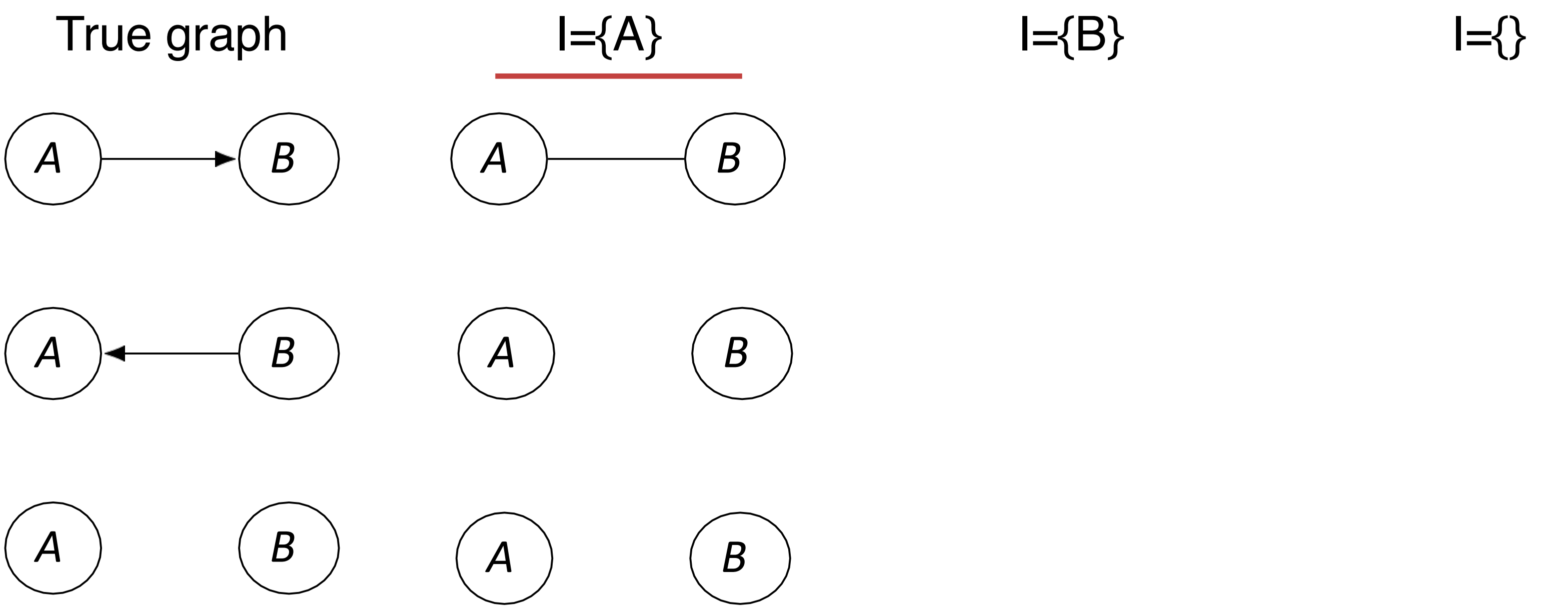
Interventional essential graphs

True graph | I={A} | I={B} | I={}



Two Variables: Interventional Essential Graphs

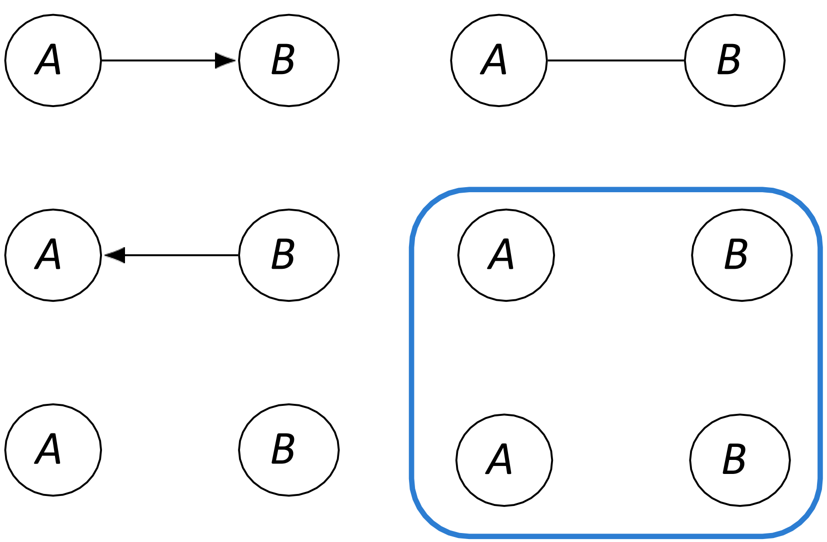
Interventional essential graphs



Two Variables: Interventional Essential Graphs

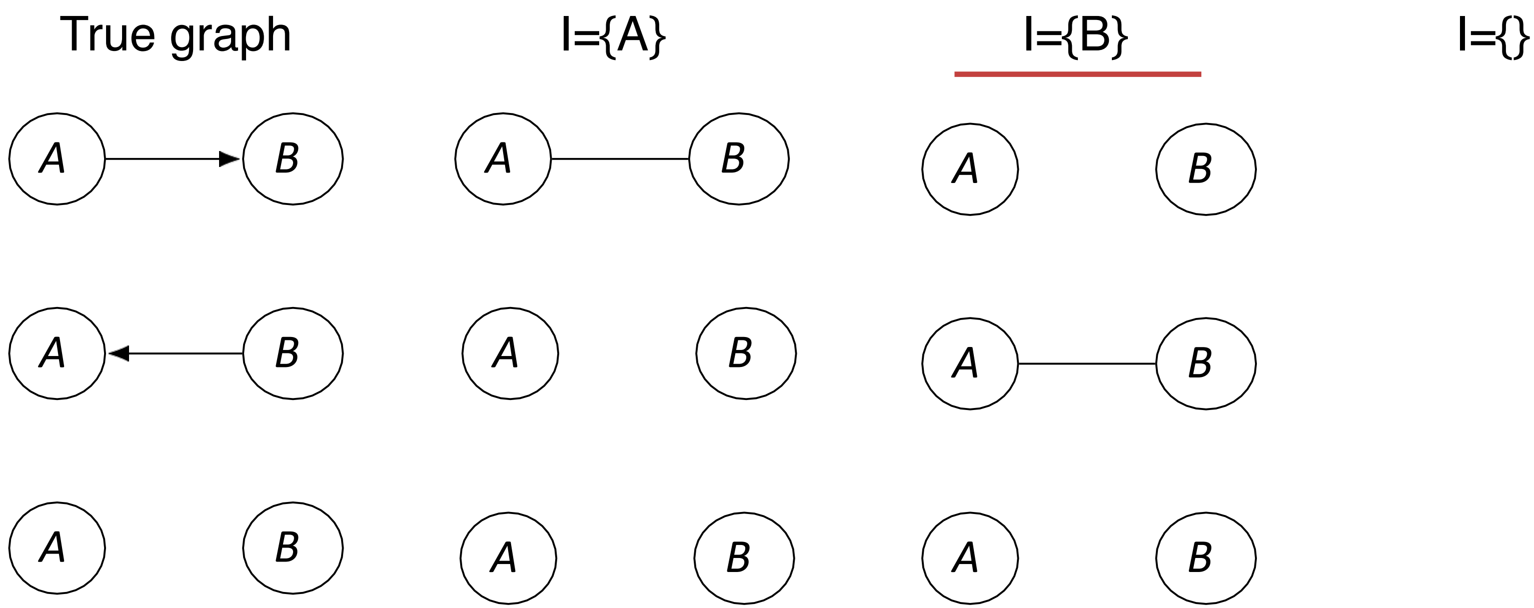
Interventional essential graphs

True graph | I={A} | I={B} | I={}

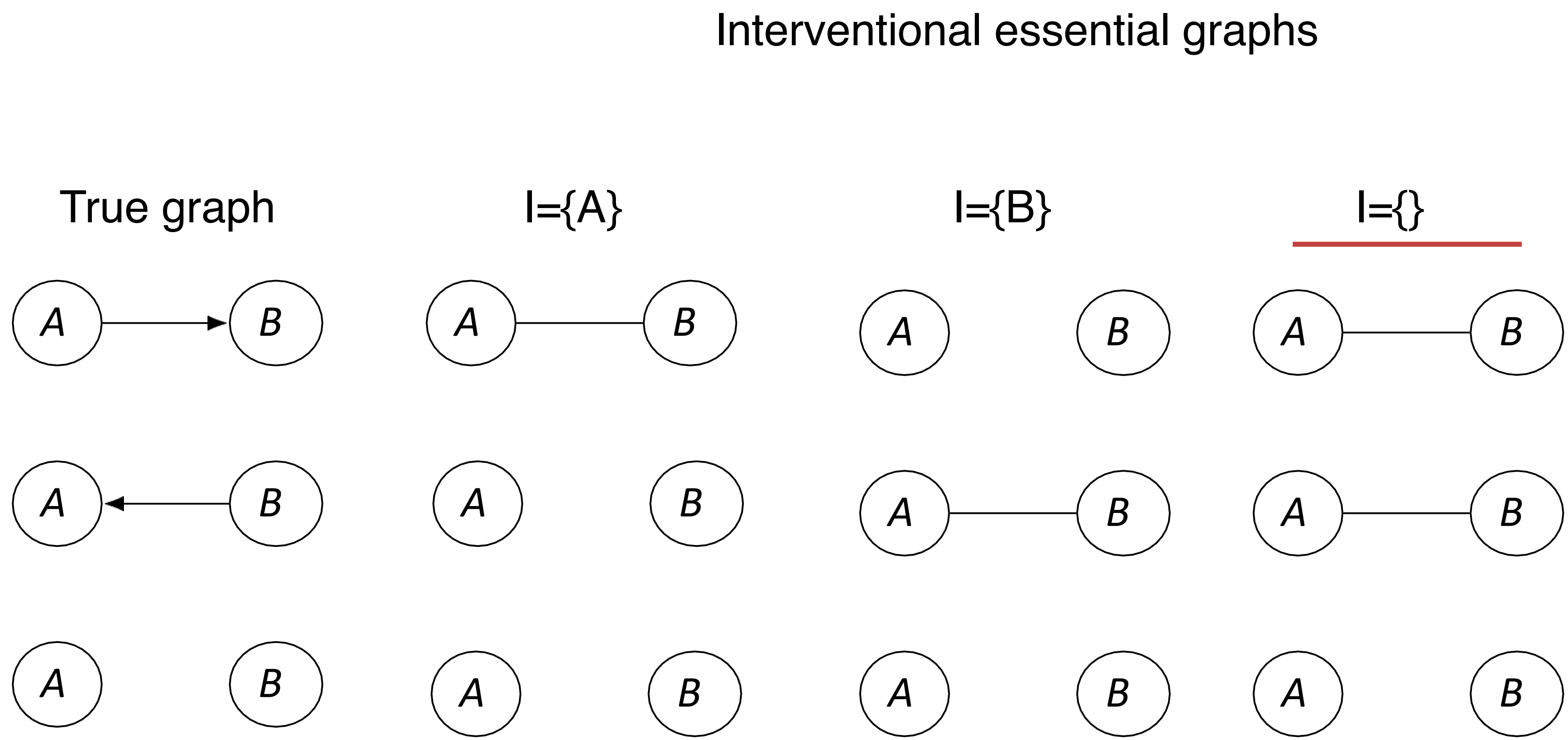


Two Variables: Interventional Essential Graphs

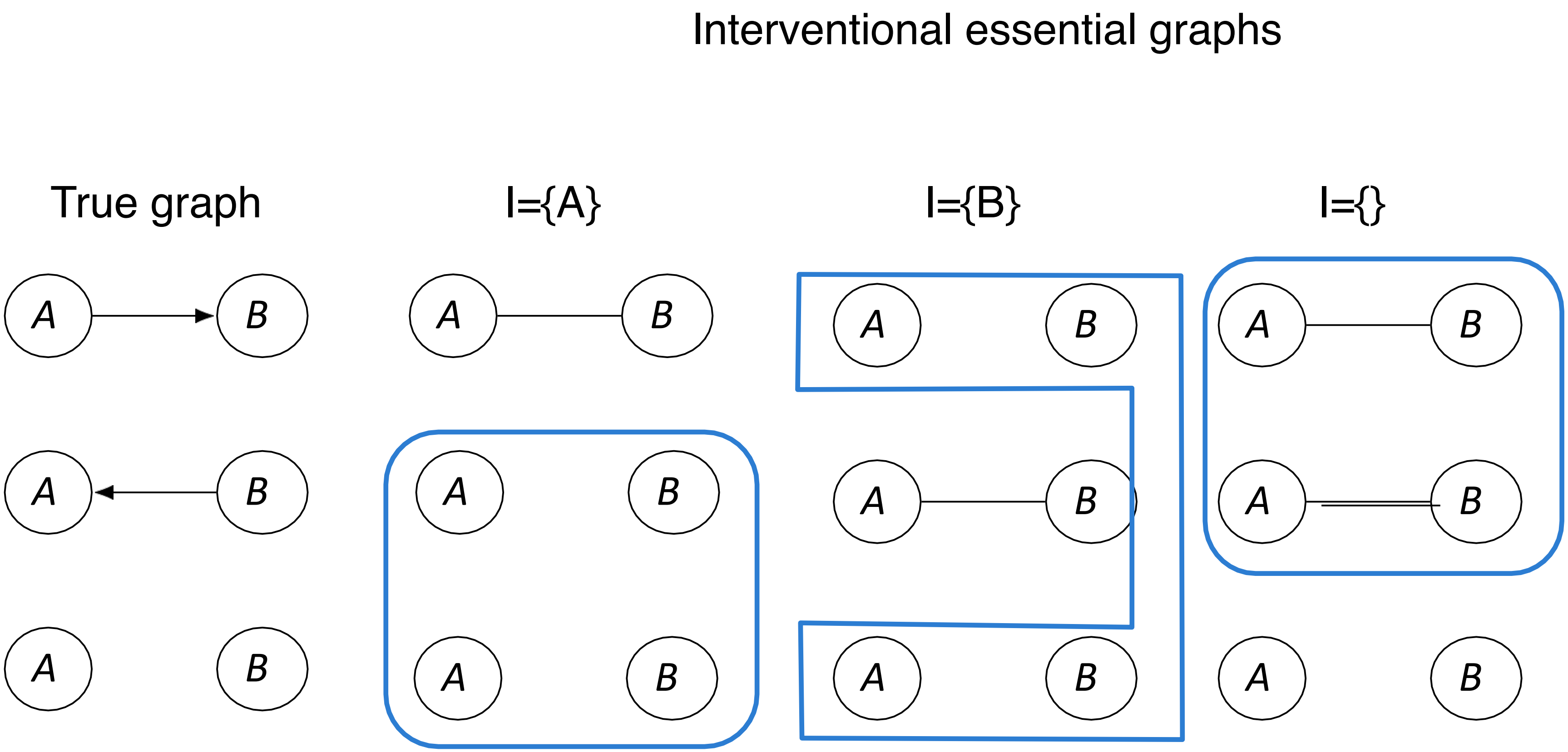
Interventional essential graphs



Two Variables: Interventional Essential Graphs



Two Variables: Interventional Essential Graphs

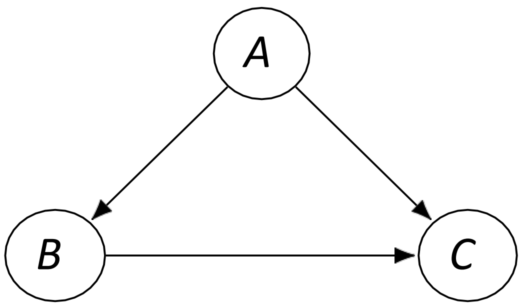


Need more than one intervention to identify the causal graph
(in this example, two interventions are sufficient and necessary to identify the graph)

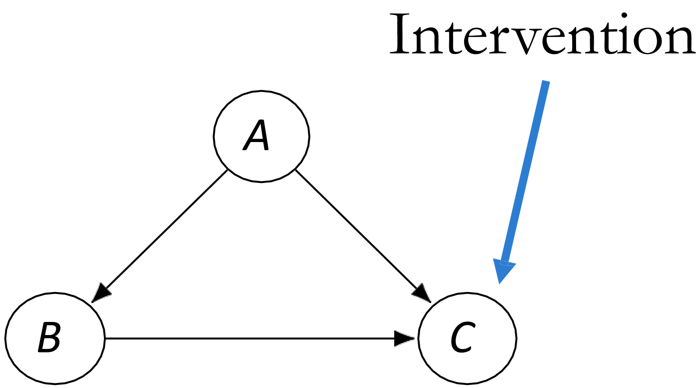
Complete Graphs Are the Worst Case

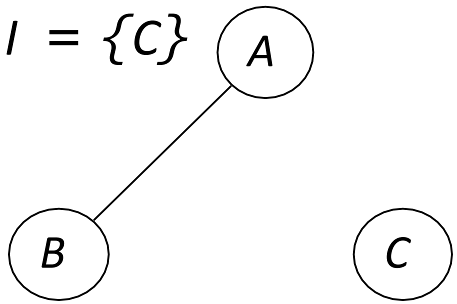
- In complete graphs, there are no immoralities, so we can only get the complete skeleton graph (e.g. from PC) as the essential graph in the worse case

Three variables



Three variables

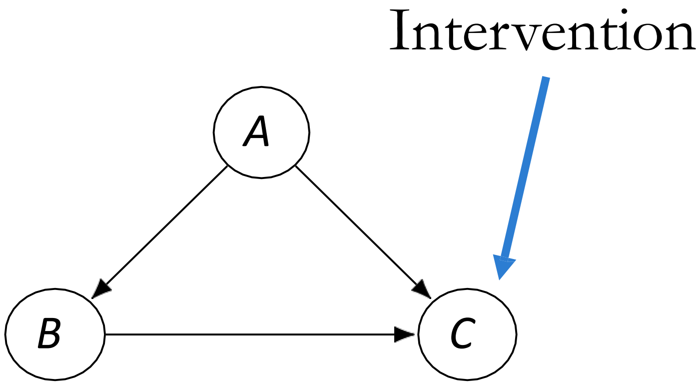




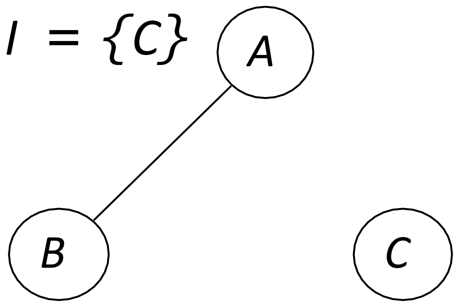
What we've learned:

- No C → A edge
- No C → B edge
- A and B connected

Three variables

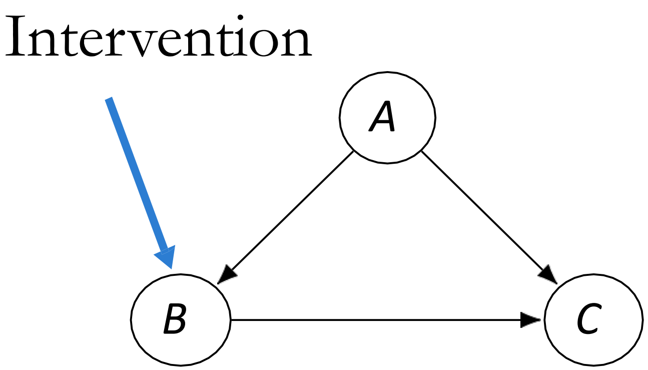


$I = \{C\}$

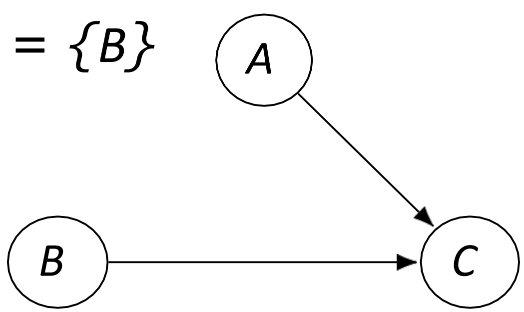


What we've learned:

- No C → A edge
- No C → B edge
- A and B connected



$I = \{B\}$



What we've learned:

- No B → A edge
- Yes A → C edge
- Yes B → C edge

When there are n variables in causal graph

- Single Variable Interventions: n – 1 interventions are sufficient for n > 2.
- In the worst case (complete graph), n – 1 single variable interventions are necessary.

Eberhardt, Frederick, Clark Glymour, and Richard Scheines. "N-1 experiments suffice to determine the causal relations among n variables." *Innovations in machine learning: theory and applications* (2006): 97-112.

Multi-Node Interventions

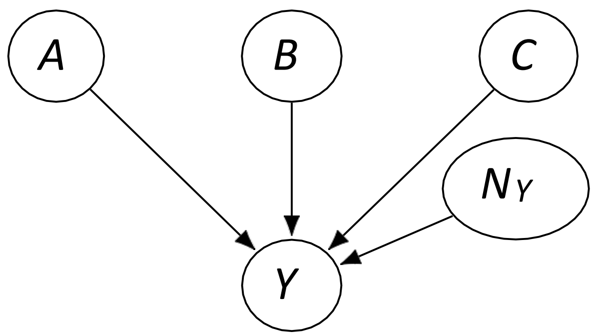
- Multi-node interventions: with no restrictions on the number of nodes per intervention
  - $[\log\_2(n)] + 1$ interventions are sufficient
  - $[\log\_2(n)] + 1$ interventions are necessary

Eberhardt, Frederick, Clark Glymour, and Richard Scheines. "On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables." *arXiv preprint arXiv:1207.1389* (2012).

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

Structural vs. Parametric Interventions

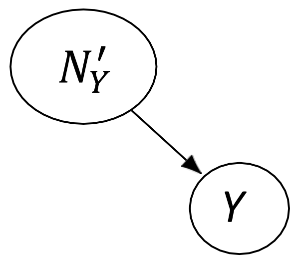


$$Y \coloneqq f_\theta(A, B, C, N_Y)$$

Structural vs. Parametric Interventions

Structural:

$Y \coloneqq N'_Y$

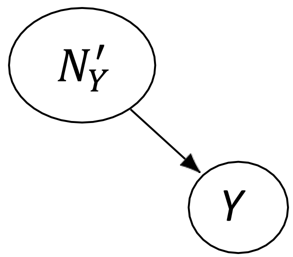


$Y \coloneqq f_\theta(A, B, C, N_Y)$

Structural vs. Parametric Interventions

Structural:

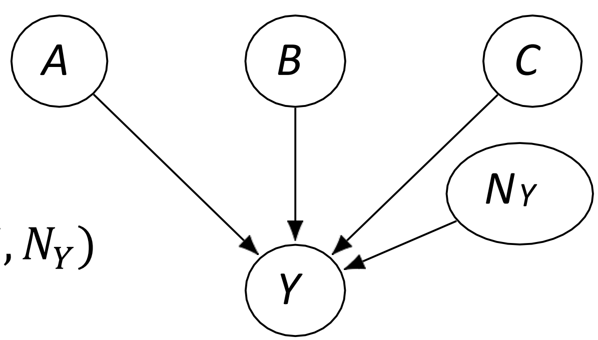
$$Y := N_Y'$$



$$Y := f_\theta(A, B, C, N_Y)$$

Parametric:

$$Y := f_{\theta'}(A, B, C, N_Y)$$

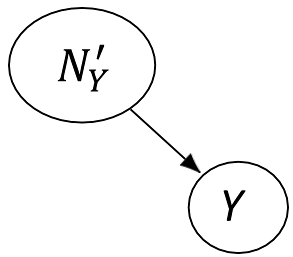


Structural vs parametric has other names: hard vs. soft, perfect vs. imperfect, etc.

Structural vs. Parametric Interventions

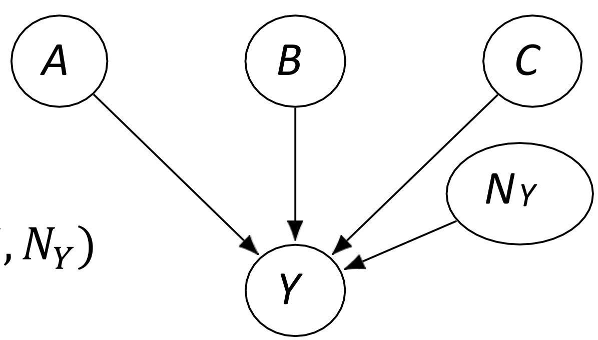
Structural:

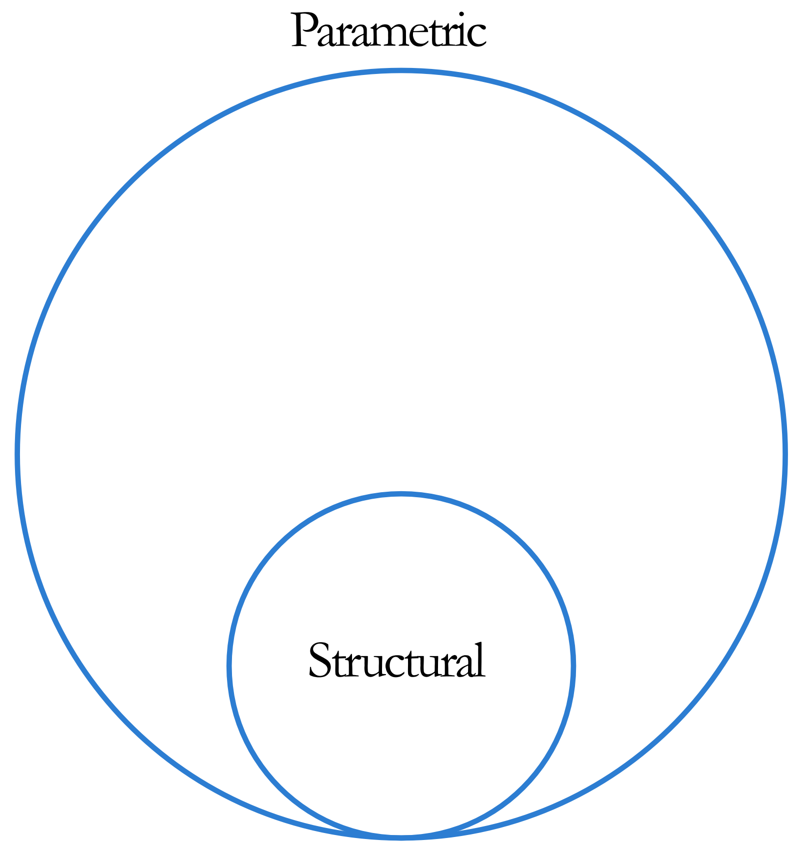
$$Y := N_Y'$$



Parametric:

$$Y := f_{\theta'}(A, B, C, N_Y)$$





Structural vs parametric has other names: hard vs. soft, perfect vs. imperfect, etc.

Number of Parametric Single-Node Interventions

- Number of interventions for identification
  - n – 1 interventions are sufficient
  - n – 1 interventions are necessary in the worst case (same as with structural interventions)

Eberhardt, Frederick, and Richard Scheines. "Interventions and causal inference." *Philosophy of science* 74.5 (2007): 981-995.

Outline

- Causal discovery from observational data
  - Independence-Based Causal Discovery
    - Assumptions
    - Markov Equivalence
    - The PC Algorithm
    - More general cases
  - Semi-Parametric Causal Discovery
    - No Identifiability Without Parametric Assumptions
    - Linear Non-Gaussian Setting
    - Nonlinear Additive Noise Setting
- Causal discovery from interventions
    - Structural intervention: Single-node / multi-node
    - Parametric intervention
    - More general cases

More General Settings

- Randomized algorithms
  - Only need $O(\log \log n)$ interventions with high probability (Hu et al., 2014) [1]
  - Better than $O(\log c)$ from Eberhardt (2008) [2] and Hauser & Bühlmann (2014) [3]
- Intervene on at most k variables per intervention: Shanmugam et al. (2015) [4]
  - Worst case: $\sim\frac{n}{2k}$
  - Randomized: $O(\frac{n}{k}\log \log k)$
- Only k interventions: Ghassami et al. (2018) [5]
- Unobserved confounding: Kocaoglu et al. (2017) [6]

[1] Hu, Huining, et al. "Randomized experimental design for causal graph discovery." *NeurIPS* 27 (2014).
[2] Eberhardt, Frederick. "Almost optimal intervention sets for causal discovery." *arXiv preprint arXiv:1206.3250* (2012).
[3] Hauser, Alain, et al. "Two optimal strategies for active learning of causal models from interventional data." *International Journal of Approximate Reasoning* 55.4 (2014).
[4] Shanmugam, Karthikeyan, et al. "Learning causal graphs with small interventions." *Advances in Neural Information Processing Systems* 28 (2015).
[5] Ghassami, AmirEmad, et al. "Budgeted experiment design for causal structure learning." *International Conference on Machine Learning*. PMLR, 2018.
[6] Kocaoglu, Murat, et al. "Experimental design for learning causal graphs with latent variables." *NeurIPS* 30 (2017).

Reading Materials

- Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models[J]. Frontiers in genetics, 2019, 10: 524.

- Eberhardt, Frederick. "Introduction to the foundations of causal discovery." *International Journal of Data Science and Analytics* 3 (2017): 81-91.

References

- Geiger D, Pearl J. On the logic of causal models[M]//Machine Intelligence and Pattern Recognition. North-Holland, 1990, 9: 3-14.
- Spirtes P, Glymour C N, Scheines R. Causation, prediction, and search[M]. MIT press, 2000.
- Brady Neal. Introduction to Causal Inference. Chapter 11
- Shimizu S, Hoyer P O, Hyvärinen A, et al. A linear non-Gaussian acyclic model for causal discovery[J]. Journal of Machine Learning Research, 2006, 7(10).
- Hoyer P O, Shimizu S, Kerminen A J, et al. Estimation of causal effects using linear non-Gaussian causal models with hidden variables[J]. International Journal of Approximate Reasoning, 2008, 49(2): 362-378.
- Lacerda G, Spirtes P L, Ramsey J, et al. Discovering cyclic causal models by independent components analysis[J]. arXiv preprint arXiv:1206.3273, 2012.

Thank you!