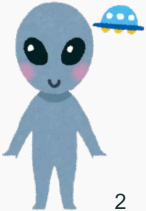
Causal Structure Discovery

Andrew Chio

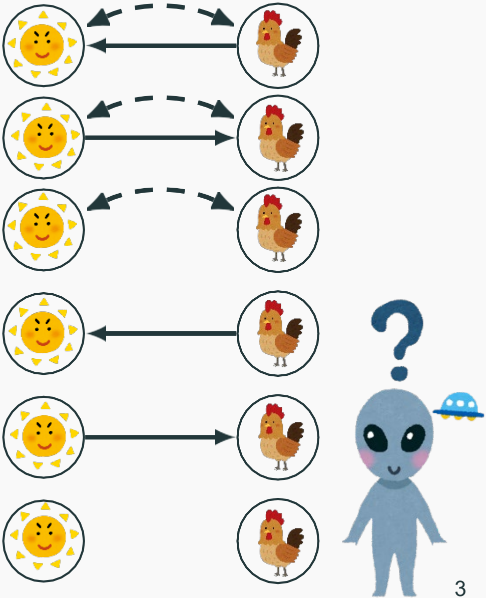
May 10, 2021

Causal Discovery: Motivation



Causal Discovery: Motivation

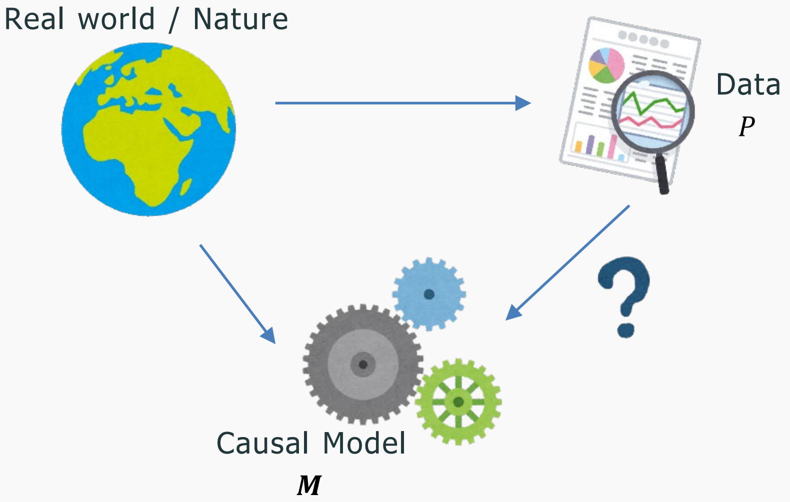
0	0
1	1
0	0
1	1
1	1
0	0
1	0
0	1
1	1



Causal Discovery

Suppose you are only given $P(V)$.

How much can you extract of the underlying causal diagram?



Review

Causal Structure of a set of variables V

A DAG where:

- Nodes = distinct element of V
- Edges = direct functional relationships between nodes

Causal Model

A 4-tuple $< V, U, \mathcal{F}, P(u) >$:

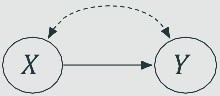
- V = endogenous variables
- U = exogenous variables
- $\mathcal{F}$ = functions which determine V :

  $$v_i \leftarrow f_i(pa_i, u_i), pa_i \subset V_i, u_i \subset U$$

- $P(u)$ = distribution over U

$$X \leftarrow f_x(U, U_x)$$

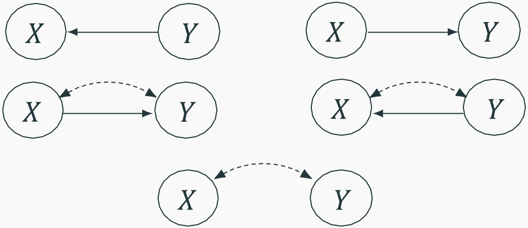
$$Y \leftarrow f_y(X, U, U_y)$$



Undirected Edges

$$Correlation \xrightarrow{?} Causal\ Structure$$

Can be either:



How can we learn causal structure?

Constraint-Based Structure Learning

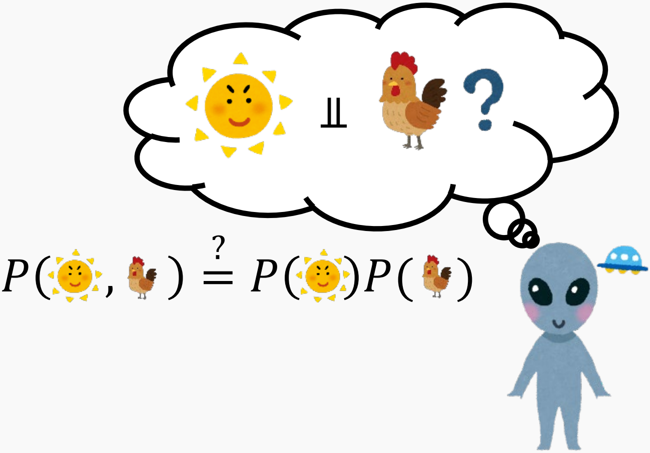
- Example
- PC & IC Algorithm
- Working with Latent Variables
- IC* Algorithm

2 other methods exist: (mentioned for completeness)

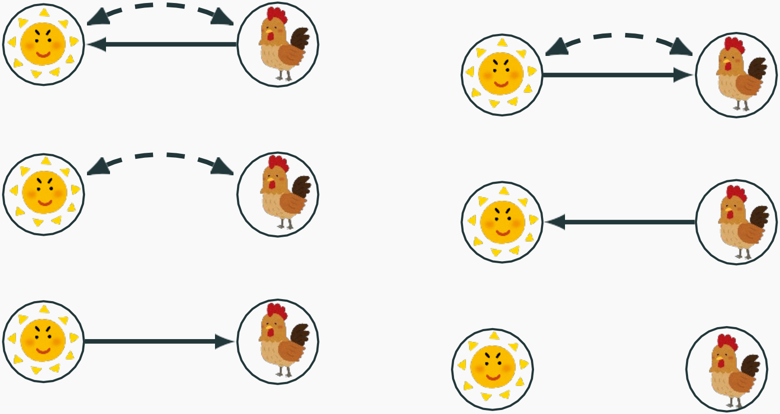
- Score-Based Structure Learning
- Function-Based Structure Learning

What constraints on the DAG exist in the data?

	
0	0
1	1
0	0
1	1
1	1
0	0
1	0
0	1
1	1

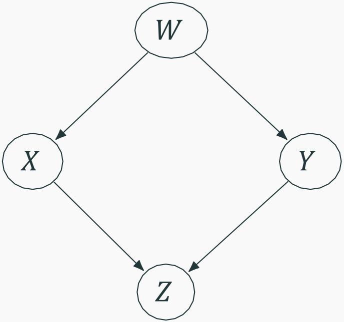


What does that mean about the graph?



What constraints does the DAG encode?

Another Example:



Independencies?

Dependencies?

$$X \perp\!\!\!\perp Y \mid W$$

$$W \perp\!\!\!\perp Z \mid XY$$

$$X \not\perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid WZ$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

The data *must* have the given independencies for this to be a compatible graph for the system.

Assumptions

Minimality [10]

If 2 graphs G_1 and G_2 can both generate $P(V)$, and G_1 can also generate any distribution G_2 generates, then G_2 is the preferred model.

Occam's razor: The most constrained model that can generate the distribution is preferred.

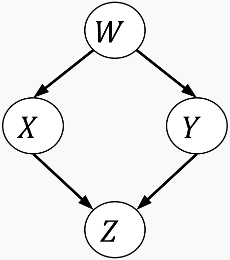
Faithfulness [12] (also called Stability [9])

The underlying natural generator does not give any independencies not immediately visible from its graphical model.

That is, if $X \perp\!\!\!\perp Y$, then the graph isn't really $X \rightarrow Y$

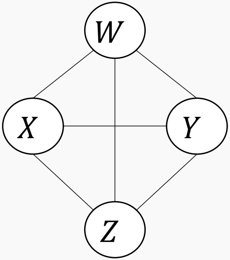
What Can We Extract?

True Model



Suppose that this graph encodes all independencies present in $P(V)$.

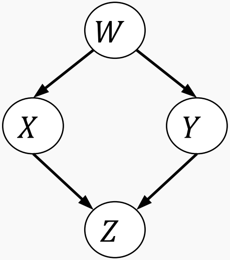
Current Best Guess



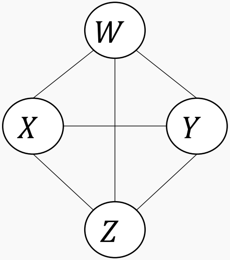
What parts of the graph can we reconstruct?

What Can We Extract?

True Model



Current Best Guess

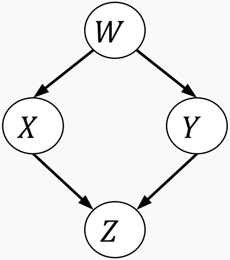


From before…

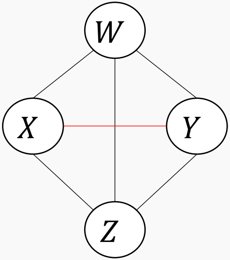
$$X \perp\!\!\!\perp Y \mid W$$
$$W \perp\!\!\!\perp Z \mid XY$$

What Can We Extract?

True Model



Current Best Guess



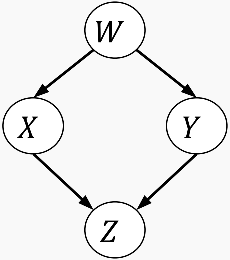
From before...

$$X \perp\!\!\!\perp Y \mid W$$

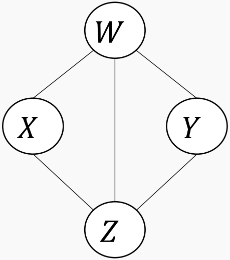
$$W \perp\!\!\!\perp Z \mid XY$$

What Can We Extract?

True Model



Current Best Guess

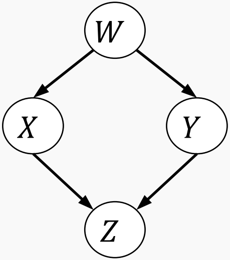


From before…

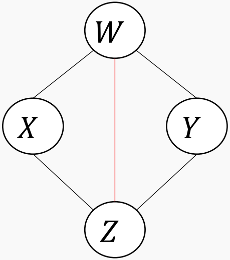
$$X \perp\!\!\!\perp Y \mid W$$
$$W \perp\!\!\!\perp Z \mid XY$$

What Can We Extract?

True Model



Current Best Guess



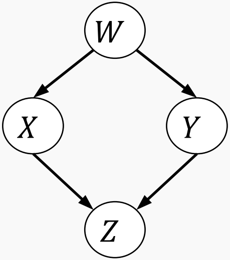
From before…

$$X \perp\!\!\!\perp Y \mid W$$

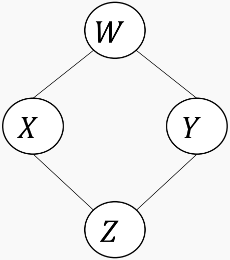
$$W \perp\!\!\!\perp Z \mid XY$$

What Can We Extract?

True Model



Current Best Guess



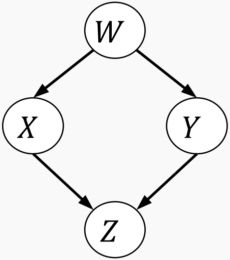
From before…

$$X \perp\!\!\!\perp Y \mid W$$
$$W \perp\!\!\!\perp Z \mid XY$$

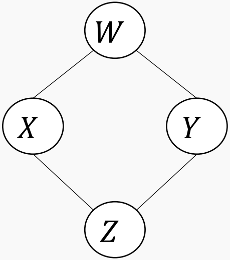
Can we reason about any edge directions?

What Can We Extract?

True Model



Current Best Guess



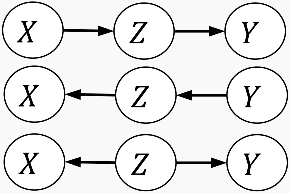
From before...

$$X \perp\!\!\!\perp Y \mid W$$

No Z!

$$W \perp\!\!\!\perp Z \mid XY$$

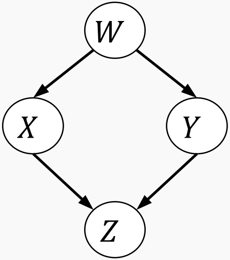
Not Possible!



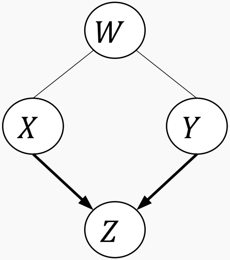
By Process of Elim:

What Can We Extract?

True Model



Current Best Guess



From before…

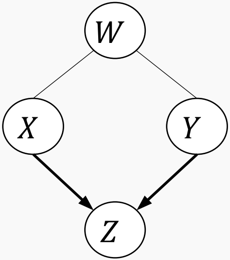
$$X \perp\!\!\!\perp Y \mid W$$
$$W \perp\!\!\!\perp Z \mid XY$$

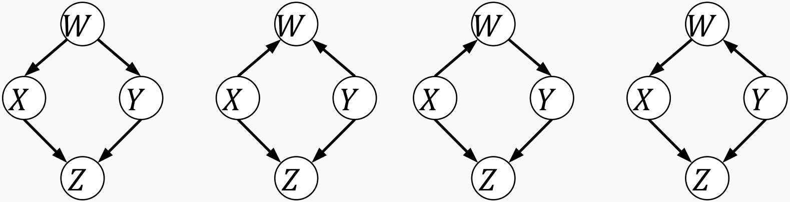
Can we do anything else?

An Equivalence Class

Equivalence Class

The set of all possible graphs that are compatible with the set of constraints that we have from the data

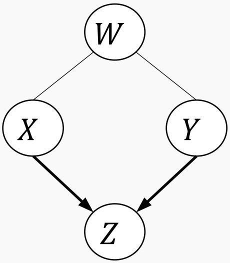




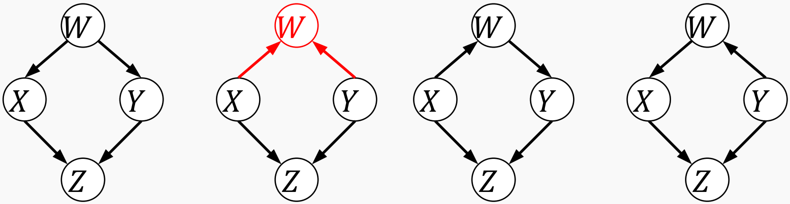
An Equivalence Class

Equivalence Class

The set of all possible graphs that are compatible with the set of constraints that we have from the data



Compatible?

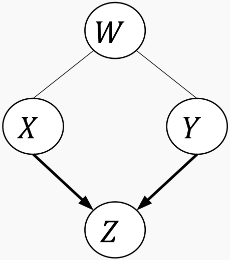


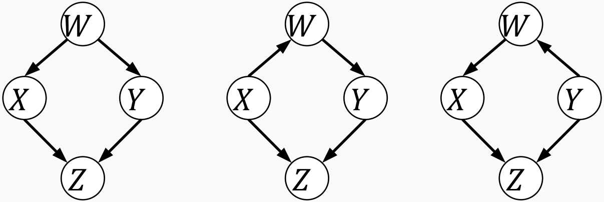
$X \perp\!\!\!\perp Y \mid W$

An Equivalence Class

Equivalence Class

The set of all possible graphs that are compatible with the set of constraints that we have from the data





PC & IC Algorithm

Assumption: True model is without latent variables and acyclic.

Input: $P(V)$

(0) Initialize empty graph G

(1) For each pair of variables $(a, b) \in V$, search for a subset of variables that makes them independent. If no such subset exists, add undirected edge $a - b$ to G

(2) For each pair of non-adjacent variables (a, b), with common neighbor c, check if c is in ab's separating set. If not, change $a - c - b$ into $a \rightarrow c \leftarrow b$

(3) In the resulting partly-directed graph, orient as many undirected edges as possible, such that:

 (a) The orientation does not add colliders that would have been found in Step 2

 (b) The orientation does not create a directed cycle

Edge Orientation Rules (for Step 3)

No New Colliders (S2), No Directed Cycles

Rules to orient edges in step 3 of previous slide:

1. Orient $b - c$ into $b \to c$ if there is $a \to b$ s.t. a, c are not adjacent.
2. Orient $a - b$ into $a \to b$ whenever there is a chain $a \to c \to b$
3. Orient $a - b$ into $a \to b$ whenever there are two chains

   $a - c \to b$ and $a - d \to b$ s.t. c, d are not adjacent
4. Orient $a - b$ into $a \to b$ whenever there are two chains

   $a - c \to d$ and $c \to d \to b$ s.t. b, c are not adjacent and

   a, d are adjacent

[IC] Reasoning for Rule 1

No New Colliders (S2), No Directed Cycles

Rule 1

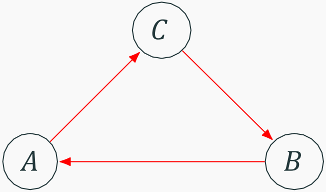
Orient $b - c$ into $b \to c$ if there is $a \to b$ s.t. a, c are not adjacent

[IC] Reasoning for Rule 2

No New Colliders (S2), No Directed Cycles

Rule 2

Orient $a - b$ into $a \to b$ whenever there is a chain $a \to c \to b$

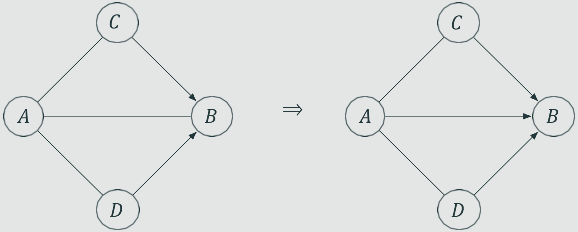


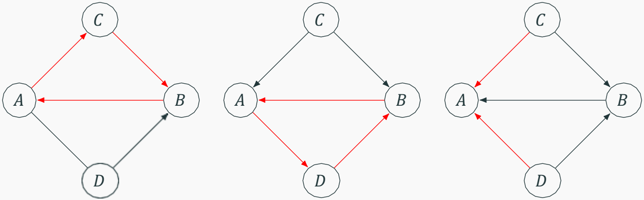
[IC] Reasoning for Rule 3

No New Colliders (S2), No Directed Cycles

Rule 3

Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow b$ and $a - d \rightarrow b$ s.t. c, d are not adjacent



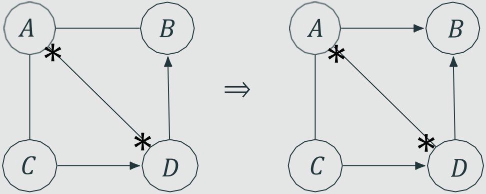


[IC] Reasoning for Rule 4

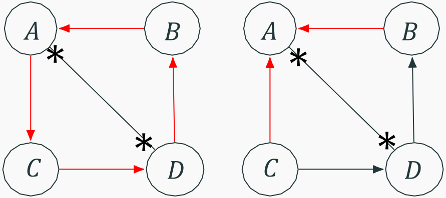
No New Colliders (S2), No Directed Cycles

Rule 4

Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow d$ and $c \rightarrow d \rightarrow b$ s.t. b, c are not adjacent and a, d are adjacent



$*-*$ represents wildcard

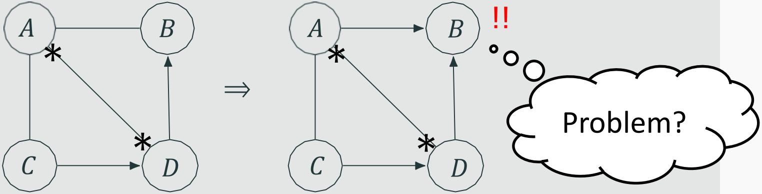


[IC] Reasoning for Rule 4

No New Colliders (S2), No Directed Cycles

Rule 4

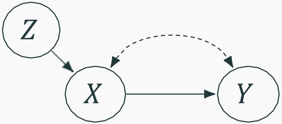
Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow d$ and $c \rightarrow d \rightarrow b$ s.t. b, c are not adjacent and a, d are adjacent

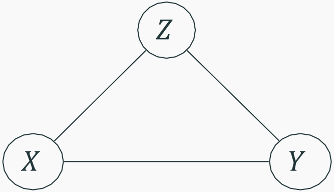


Doesn't matter that B is a collider; A, D are already dependent

Dealing with Latents

What happens if we run IC on a model with latent variables?





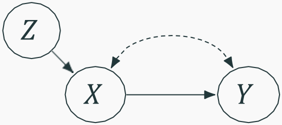
The edges do not represent direct causation anymore!

PDAGs: Partial DAGs

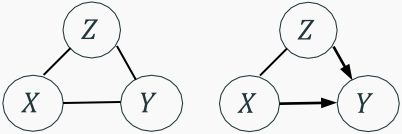
PDAG

A DAG representing incomplete information about the underlying causal model. It has several types of edges:

1. Marked arrow $a \overset{*}{\rightarrow} b$ signifies a directed path a to b
2. Unmarked arrow $a \rightarrow b$ signifies either a directed path or a latent variable (or both)
3. Bidirected edge $a \leftrightarrow b$ signifies a latent common cause
4. An undirected edge $a - b$ signifies a latent variable, $a \rightarrow b$, or $a \leftarrow b$



True Model



Compatible PDAGs

IC*

(0) Initialize empty graph G

(1) For each pair of variables $(a, b) \in V$, search for a subset of variables that makes them independent. If no such subset exists, add undirected edge $a - b$ to G **[Same as IC]**

(2) For each pair of non-adjacent variables (a, b), with common neighbor c, check if c is in ab's separating set. If not, change $a - c - b$ into $a \rightarrow c \leftarrow b$ **[Same as IC]**

(3) In the resulting PDAG, add as many arrowheads as possible, and mark as many edges as possible, according to:

   (a) Orient $b -\!* c$ into $b \rightarrow c$ if there is $a *\!\!\rightarrow b$ s.t. a, c are not adjacent

   (b) If a, b are adjacent and there is a directed path from a to b, then set $a * -b$ to $a *\!\!\rightarrow b$

Note on Notation: Overloaded *

Edges with * above them

Represents a directed path

e.g., $a \overset{*}{\rightarrow} b$

Edges with * at end

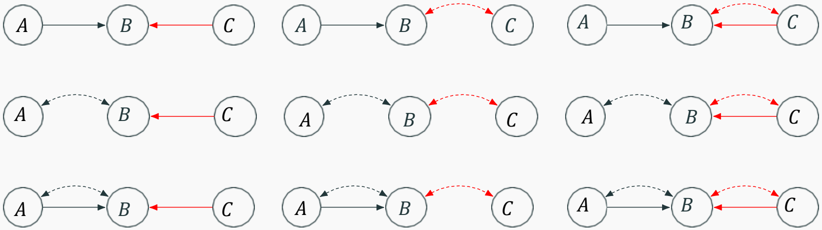
Represents a wildcard (we do not care what arrow is there)

e.g., $a \ast\!\!\rightarrow b$ can be $a \leftrightarrow b$ or $a \rightarrow b$

[IC*] Reasoning on Rule 1

Rule 1

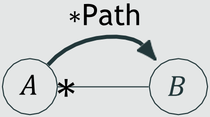
Orient $b -\!* \; c$ into $b \overset{*}{\rightarrow} c$ if there is $a * \!- b$ s.t. a, c are not adjacent



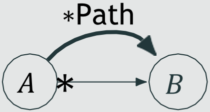
[IC*] Reasoning on Rule 2

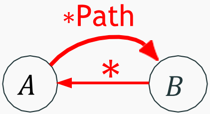
Rule 2

If a, b are adjacent and there is a directed path from a to b using only edges $\overset{*}{\rightarrow}$, then set $a * \!- b$ to $a *\!\!\rightarrow b$



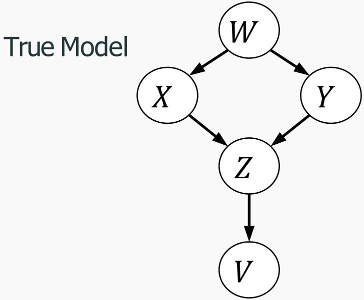
$\Rightarrow$

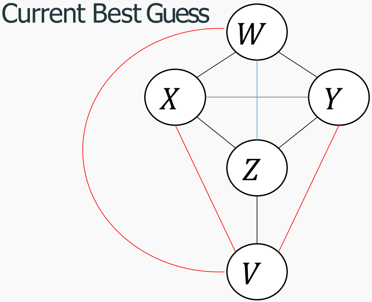




Adding the arrowhead only disallows this graph all others are still allowed.

IC* Example





Start as before:

1. Eliminate edges between d-separated nodes

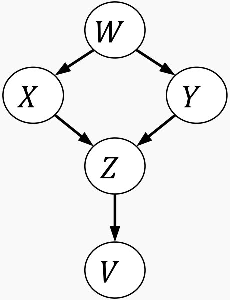
$X \perp\!\!\!\perp Y \mid W$

$W \perp\!\!\!\perp Z \mid XY$

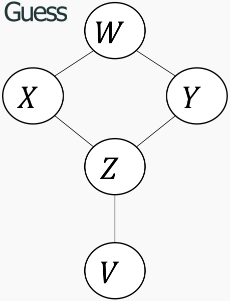
$WXY \perp\!\!\!\perp V \mid Z$

IC* Example

True Model



Current Best Guess



Start as before:

1. Eliminate edges between d-separated nodes

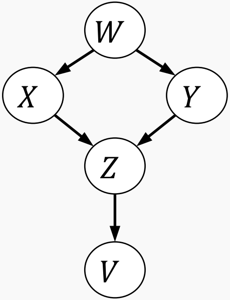
$X \perp\!\!\!\perp Y \mid W$

$W \perp\!\!\!\perp Z \mid XY$

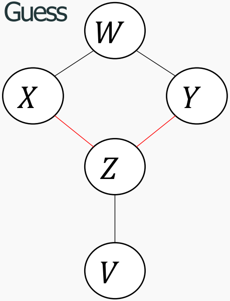
$WXY \perp\!\!\!\perp V \mid Z$

IC* Example

True Model



Current Best Guess



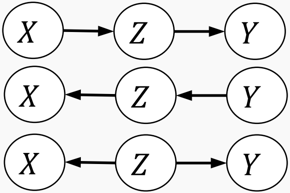
Start as before:

2. Orient discoverable colliders

$X \perp\!\!\!\perp Y \mid W$

No Z!

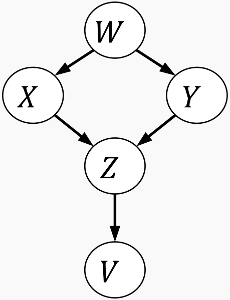
Not Possible!



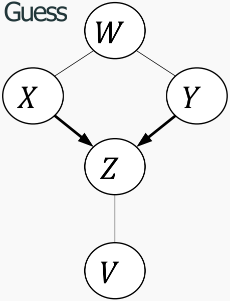
By Process of Elim:

IC* Example

True Model



Current Best Guess

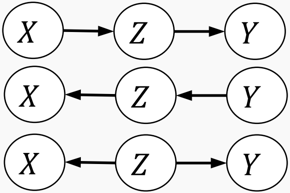


Start as before:

2. Orient discoverable colliders

$X \perp\!\!\!\perp Y \mid W$ No Z!

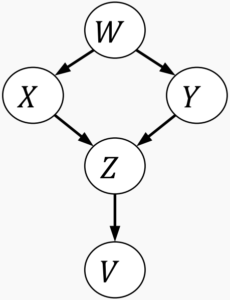
Not Possible!



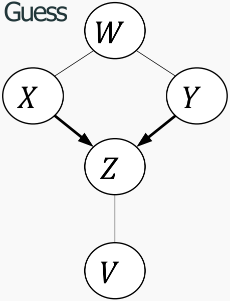
By Process of Elim:

IC* Example

True Model



Current Best Guess



[IC*] Can we apply any rules?

Rule 1

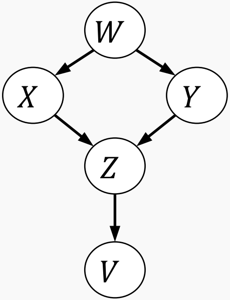
Orient $b - * c$ into $b \overset{*}{\rightarrow} c$ if there is $a * - b$ s.t. a, c are not adjacent

Rule 2

If a, b are adjacent and there is a directed path from a to b using only edges $\overset{*}{\rightarrow}$, then set $a * - b$ to $a * \rightarrow b$

IC* Example

True Model

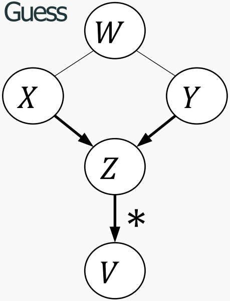


[IC*] Rule 1:

$Z -\!* V$ to $Z \overset{*}{\to} V$ since

$X *\!- Z$ and X, V are not adj.

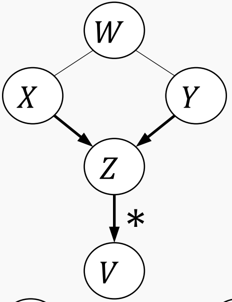
Current Best Guess



Anything else?

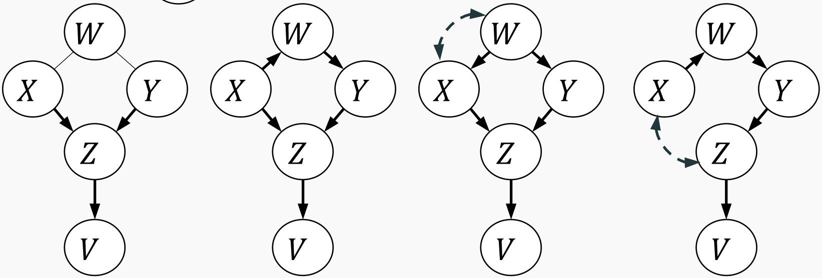
IC* Example

Equivalence Class



PDAG Arrows

- $a \overset{*}{\rightarrow} b$: directed path a to b
- $a \rightarrow b$: directed path and/or latent variable
- $a \leftrightarrow b$: a latent common cause
- $a - b$: a latent variable, $a \rightarrow b$, or $a \leftarrow b$



The constraint-based approach to determining $x - y$

- Sometimes, we only care about determining causal relationship between X, Y
- Steps:
  - Check if $X \perp\!\!\!\perp Y$
  - If not, find other variables in the system correlated with X, Y.
  - Repeat* until learned graph can allow you to orient edge $X - Y$, or no possible sources of data remain

* Using a similar algorithm known as FCI [13], which was shown to be complete for edge orientation [14] and utilizes a different encoding of graph called PAG.

Summary

- Conditional Independence Constraints allow us to extract partial information about underlying graphical structure
  - … but they are not always sufficient to extract the full graph
- Recent Research has extended notions into PAGs (e.g., identifiability) [4]

References

[1] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, volume 3734, pages 63–77. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-29242-5 978-3-540-31696-1. doi: 10.1007/11564089 7.

[2] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc., 2009.

[3] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4):411–430, June 2000. ISSN 0893-6080. doi: 10.1016/S0893-6080(00)00026-5.

References

[4] A. Jaber, J. Zhang, and E. Bareinboim. Causal Identification under Markov Equivalence. 2018.

[5] D. Janzing and B. Schoelkopf. Causal inference using the algorithmic Markov condition. *arXiv:0804.3678 [cs, math, stat]*, Apr. 2008.

[6] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniuˇsis, B. Steudel, and B. Sch¨olkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, May 2012. ISSN 00043702. doi: 10.1016/j.artint.2012.01.002.

[7] R. Jiao, N. Lin, Z. Hu, D. A. Bennett, L. Jin, and M. Xiong. Bivariate Causal Discovery and Its Applications to Gene Expression and Imaging Data Analysis. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00347.

[8] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Scholkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. page 102.

[9] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. ISBN 978-0-521-89560-6.

[10] J. Pearl and T. S. Verma. A theory of inferred causation. In D. Prawitz, B. Skyrms, and D. Westerståhl, editors, *Studies in Logic and the Foundations of Mathematics*, volume 134 of *Logic, Methodology and Philosophy of Science IX*, pages 789–811. Elsevier, Jan. 1995. doi: 10.1016/S0049-237X(06)80074-1.

References

[11] S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. page 28, 2006.

[12] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, Prediction, and Search*. MIT Press, 1993. ISBN 978-0-262-19440-2.

[13] J. Zhang. A Characterization of Markov Equivalence Classes for Directed Acyclic Graphs with Latent Variables. page 8, 2007.

[14] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172 (16-17):1873–1896, Nov. 2008. ISSN 00043702. doi: 10.1016/j.artint.2008.08.001.