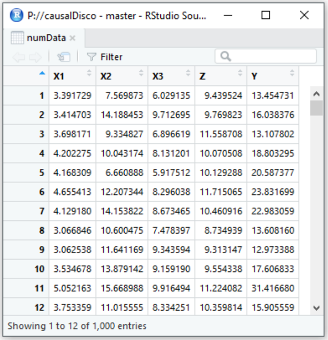
UNIVERSITY OF COPENHAGEN

Introduction to causal discovery: CPDAGs and the PC algorithm

Anne Helby Petersen

A statistician's dream

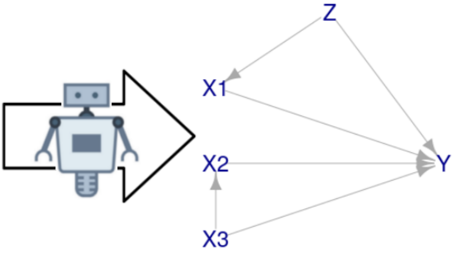

P://causalDisco - master - RStudio Sou...

numData

Filter

	X1	X2	X3	Z	Y
1	3.391729	7.569873	6.029135	9.439524	13.454731
2	3.414703	14.188453	9.712695	9.769823	16.038376
3	3.698171	9.334827	6.896619	11.558708	13.107802
4	4.202275	10.043174	8.131201	10.070508	18.803295
5	4.168309	6.660888	5.917512	10.129288	20.587377
6	4.655413	12.207344	8.296038	11.715065	23.831699
7	4.129180	14.153822	8.673465	10.460916	22.983059
8	3.066846	10.600475	7.478397	8.734939	13.608160
9	3.062538	11.641169	9.343594	9.313147	12.973388
10	3.534678	13.879142	9.159190	9.554338	17.606833
11	5.052163	15.668988	9.916494	11.224082	31.416680
12	3.753359	11.015555	8.334251	10.359814	15.905559

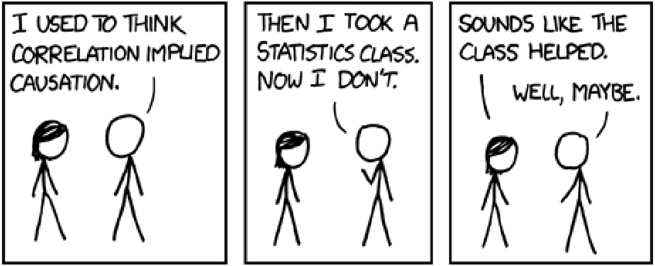
Showing 1 to 12 of 1,000 entries



Why it would be great

- Constructing DAGs is time consuming and difficult
- Risk of confirmation bias when basing causal inference on "expert-made" DAG: We can only find what we are looking for
- Different experts end up making different DAGs $\Rightarrow$ current standard approach is not ideal

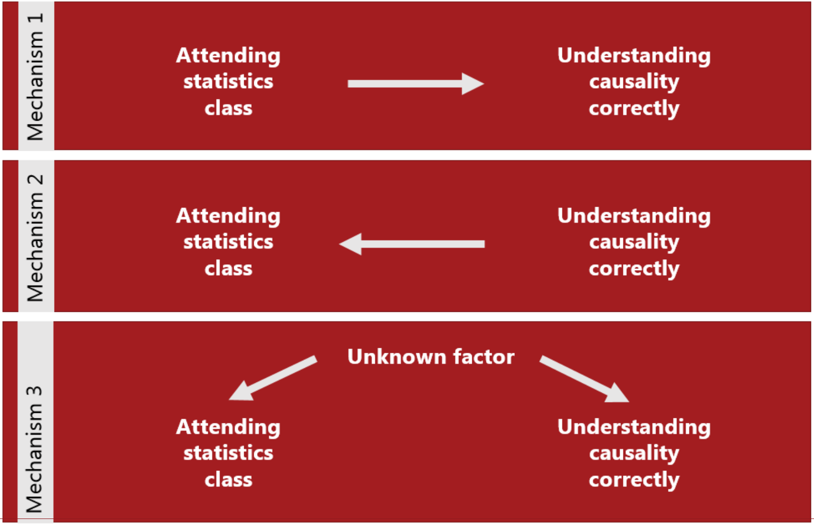
Correlation does **not** imply causation



Source: www.xkcd.com/552/

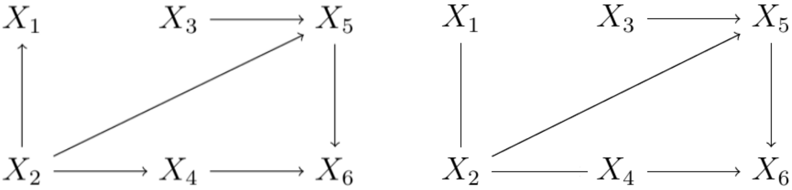
… but causation may imply association

Reichenbach's common cause principle: An association occurs due to one of three possible mechanisms:



DAGs and CPDAGs

Directed acyclic graphs and completed partially directed acyclic graphs



- **DAG interpretation:** Directed edge from X to Y means that X is a direct cause of Y.
- **Markov property:** DAG structure (d-separations) $\Rightarrow$ conditional independencies in distribution.
- A CPDAG describes a Markov **equivalence class**, i.e., the set of all DAGs that imply the same conditional independencies.
- **CPDAG interpretation:** Undirected edges denotes ambivalence about edge orientation within equivalence class. Directed edges are interpreted as for DAGs.

Causal assumptions

No free lunch, need to make some untestable assumptions:

1. Faithfulness: Conditional independencies in distribution $\Rightarrow$ DAG structure (d-separations)
   (reverse implication of Markov property)
2. Acyclic data generating mechanisms: No feedback loops
3. No conditioning on unobserved colliders
4. No unobserved confounding

Causal assumptions

No free lunch, need to make some untestable assumptions:

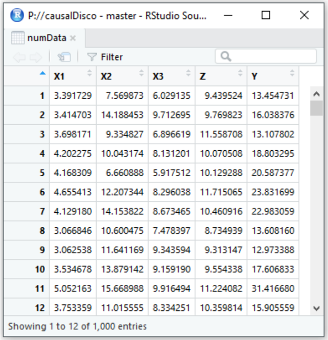
1. Faithfulness: Conditional independencies in distribution $\Rightarrow$ DAG structure (d-separations)
(reverse implication of Markov property)
2. Acyclic data generating mechanisms: No feedback loops
3. No conditioning on unobserved colliders
4. No unobserved confounding

Causal assumptions

No free lunch, need to make some untestable assumptions:

1. Faithfulness: Conditional independencies in distribution $\Rightarrow$ DAG structure (d-separations)
   (reverse implication of Markov property)
2. Acyclic data generating mechanisms: No feedback loops
3. No conditioning on unobserved colliders
4. No unobserved confounding - relaxed later today!

A statistician's dream version 2.0



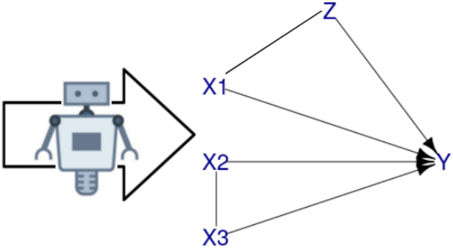
P://causalDisco - master - RStudio Sou...

numData

Filter

	X1	X2	X3	Z	Y
1	3.391729	7.569873	6.029135	9.439524	13.454731
2	3.414703	14.188453	9.712695	9.769823	16.038376
3	3.698171	9.334827	6.896619	11.558708	13.107802
4	4.202275	10.043174	8.131201	10.070508	18.803295
5	4.168309	6.660888	5.917512	10.129288	20.587377
6	4.655413	12.207344	8.296038	11.715065	23.831699
7	4.129180	14.153822	8.673465	10.460916	22.983059
8	3.066846	10.600475	7.478397	8.734939	13.608160
9	3.062538	11.641169	9.343594	9.313147	12.973388
10	3.534678	13.879142	9.159190	9.554338	17.606833
11	5.052163	15.668988	9.916494	11.224082	31.416680
12	3.753359	11.015555	8.334251	10.359814	15.905559

Showing 1 to 12 of 1,000 entries



Goal: Estimate CPDAG by analyzing data (i.e., causal discovery).

Overall idea: Causal relationships leave behind traces in data (conditional independencies) that can be used to reconstruct parts of the causal model (its Markov equivalence class).

Focus of today: Causal discovery algorithms making use of conditional independence testing (constraint-based).

Ingredients for a causal discovery algorithm

Recall:

- d-separation: Two variables X and Y are d-separated by at set of variables $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ if the following two conditions hold:
  1. All causal paths ($X \to \ldots \to Z_i \to \ldots \to T$ or confounder paths ($X \leftarrow \ldots \leftarrow Z_i \to \ldots \to Y$) between X and Y include a variable Z_i from $\mathbf{Z}$.
  2. No collider paths ($X \to \ldots \to Z_i \leftarrow \ldots \leftarrow Y$) between X and Y include a variable Z_i from $\mathbf{Z}$, nor a descendant of any variable in $\mathbf{Z}$.
- Assuming the Markov property and faithfulnes, we get that X and Y are d-separated by Z exactly when X and Y are conditionally independent given Z.
- Note: We can test conditional independence using empirical data!

The PC algorithm (Spirtes & Glymour 1991)

Peter-Clark (PC) algorithm summary

Input: Information about conditional independencies[a]

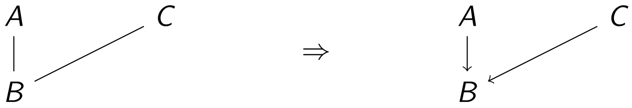
1. Start with fully connected undirected graph
2. Repeat: For each pair of variables (A, B), look for separating sets S among variables adjancent to A or B s.t. $A \perp\!\!\!\perp B \mid S$. If such an S exists: Remove edge between A and B.
3. Apply orientation rules making use of v-structures and acyclicity assumption

Output: CPDAG

[a] In practice we use statistical tests to determine conditional independence.

PC orientation rules

First, apply **v-structure orientation**: For each structure $A - B - C, A \not- C$: orient as $A \rightarrow B \leftarrow C$ if $B \notin S$ for all $\mathbf{S}$ such that $A \perp\!\!\!\perp C \mid \mathbf{S}$.



Next, recursively apply **three additional rules** (next slide) until no further changes are made.

These rules are **sound and complete** (in the large sample limit): No incorrect orientations occur, and no further orientations can be made (Meek 1995).

Meek's orientation rules

R1: Avoid introducing new v-structures (directly):

$$
\begin{tikzcd}
A \arrow[d] & C \\
B \arrow[ur, dash] &
\end{tikzcd}
\quad \Rightarrow \quad
\begin{tikzcd}
A \arrow[d] & C \\
B \arrow[ur] &
\end{tikzcd}
$$

R2: Avoid introducing cycles.

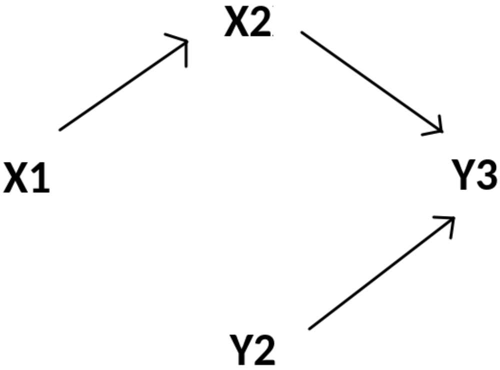
$$
\begin{tikzcd}
A \arrow[d] \arrow[r, dash] & C \\
B \arrow[ur] &
\end{tikzcd}
\quad \Rightarrow \quad
\begin{tikzcd}
A \arrow[d] \arrow[r] & C \\
B \arrow[ur] &
\end{tikzcd}
$$

R3: Avoid introducing new v-structures (indirectly).

$$
\begin{tikzcd}
A \arrow[r, dash] \arrow[d, dash] \arrow[dr, dash] & C \arrow[d] \\
B \arrow[r] & D
\end{tikzcd}
\quad \Rightarrow \quad
\begin{tikzcd}
A \arrow[r, dash] \arrow[d, dash] \arrow[dr] & C \arrow[d] \\
B \arrow[r] & D
\end{tikzcd}
$$

PC algorithm example

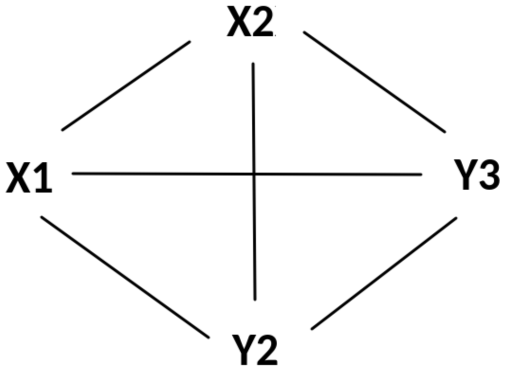
True graph:



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

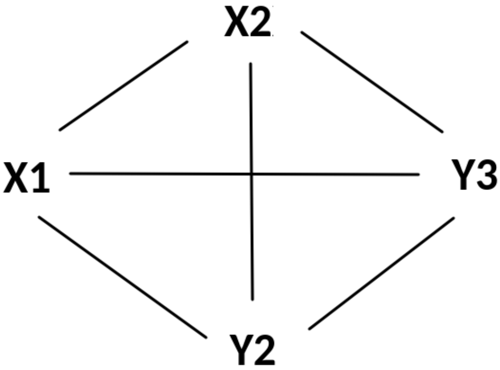
Start with fully connected graph



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

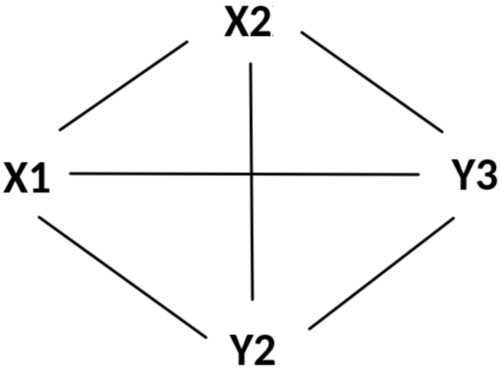
For each pair of adjacent variables, look for separating sets of size 0



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

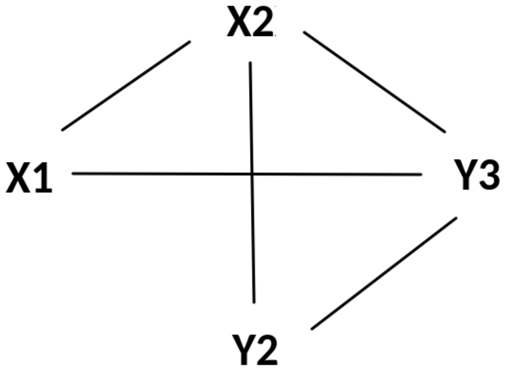
For each pair of adjacent variables, look for separating sets of size 0



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

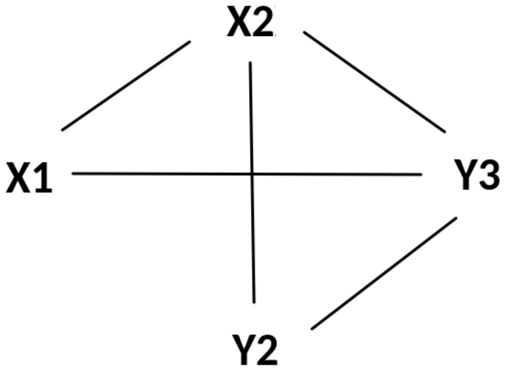
For each pair of adjacent variables, look for separating sets of size 0



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

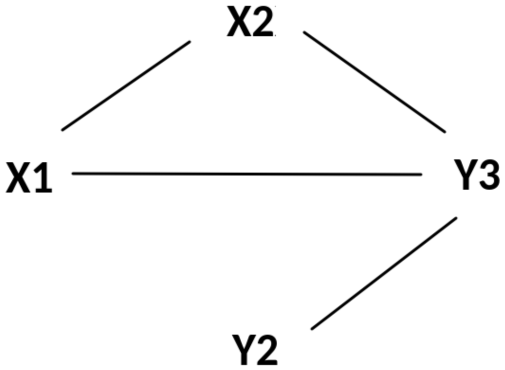
For each pair of adjacent variables, look for separating sets of size 0



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

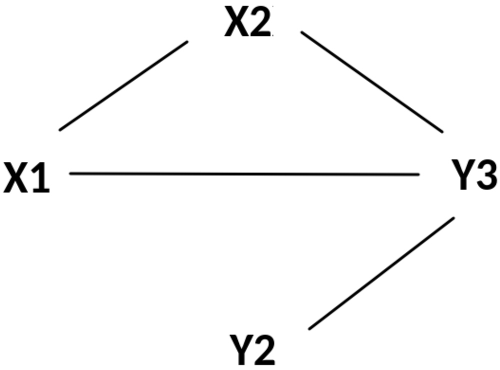
For each pair of adjacent variables, look for separating sets of size 0



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3|X2$.

PC algorithm example

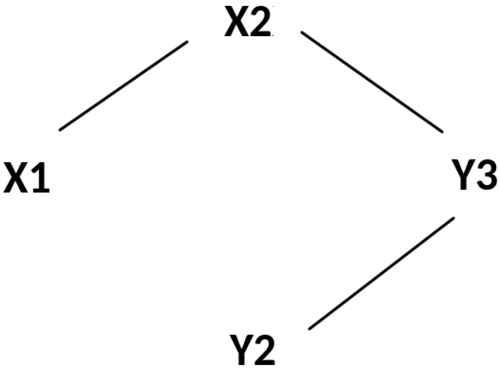
For each pair of adjacent variables, look for separating sets of size 1



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3|X2$.

PC algorithm example

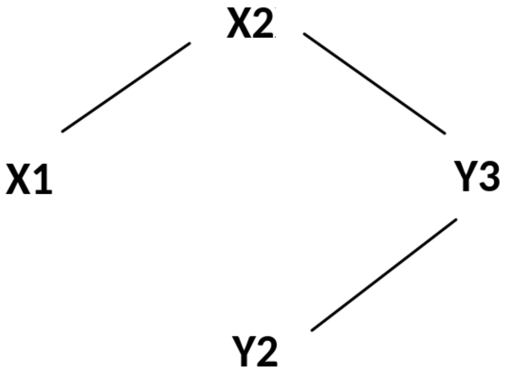
For each pair of adjacent variables, look for separating sets of size 1



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

PC algorithm example

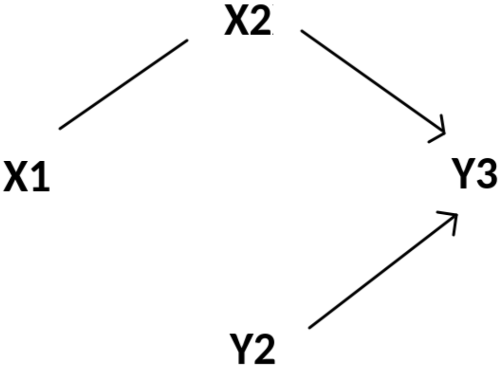
Orient v-structures



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3|X2$.

PC algorithm example

Orient v-structures - $Y3 \notin S$ for any S s.t. $X2 \perp\!\!\!\perp Y2 | S$.



Cond. indep.: $X1 \perp\!\!\!\perp Y2$, $X2 \perp\!\!\!\perp Y2$, $X1 \perp\!\!\!\perp Y3 | X2$.

Choices to be made

Using PC on empirical data requires one to choose:

1. A conditional independence test.
2. A significance level to use in the tests.

Choices to be made

Using PC on empirical data requires one to choose:

1. A conditional independence test.
2. A significance level to use in the tests.

Note:

- There does not exist a generally correct tests of conditional independence which does not rely on some distributional assumptions (Shah & Petersen 2020).
- We do not have a principled approach for choosing the test level.

Conditional independence testing

We do have some simple examples where correct tests[1] do exist:

[1] Up to statistical uncertainty...

Conditional independence testing

We do have some simple examples where correct tests[1] do exist:

If the data are **jointly normally distributed**, we have:

$$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow \mathrm{cor}(X, Y \mid Z) = 0$$

Note that $\mathrm{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta = 0$ in the linear regression model

$$Y_i = \alpha + \beta \cdot X_i + \gamma \cdot Z_i + \epsilon_i$$

[1] Up to statistical uncertainty...

Conditional independence testing

We do have some simple examples where correct tests[1] do exist:

If the data are **jointly normally distributed**, we have:

$$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow \mathrm{cor}(X, Y \mid Z) = 0$$

Note that $\mathrm{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta = 0$ in the linear regression model

$$Y_i = \alpha + \beta \cdot X_i + \gamma \cdot Z_i + \epsilon_i$$

If the data are **exclusively categorical**, we can directly test conditional independence by use of e.g. a χ^2 test of independence on the multiway cross tabulation over X, Y, Z.

[1] Up to statistical uncertainty...

Conditional independence testing

We do have some simple examples where correct tests[1] do exist:

If the data are **jointly normally distributed**, we have:

$$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow \mathrm{cor}(X, Y \mid Z) = 0$$

Note that $\mathrm{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta = 0$ in the linear regression model

$$Y_i = \alpha + \beta \cdot X_i + \gamma \cdot Z_i + \epsilon_i$$

If the data are **exclusively categorical**, we can directly test conditional independence by use of e.g. a χ^2 test of independence on the multiway cross tabulation over X, Y, Z.

Today, we will (pragmatically) test a necessary condition for conditional independence for mix of binary/numeric variables: Test for non-association using **GLMs with spline-expansions** (Petersen, Osler & Ekstrøm 2021).

[1] Up to statistical uncertainty...

Test level

- The significance level used for individual tests in the PC algorithm is *not* a proper significance level for the globally estimated graph
  - It does not describe the overall risk of type I error
  - Many tests are conducted, and the result of one test informs what test should be conducted next $\Rightarrow$ a complicated multiple testing issue without obvious solutions

Test level

- The significance level used for individual tests in the PC algorithm is *not* a proper significance level for the globally estimated graph
  - It does not describe the overall risk of type I error
  - Many tests are conducted, and the result of one test informs what test should be conducted next $\Rightarrow$ a complicated multiple testing issue without obvious solutions
- Today, we will pragmatically consider an arbitrary choice of $\alpha = 0.05$ (exercises regarding varying this).

References

Meek (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95*.

Petersen, Osler & Ekstrøm (2021). Data-driven model building for life-course epidemiology. *American Journal of Epidemiology*.

Shah & Peters (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*.

Spirtes & Glymour (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*.