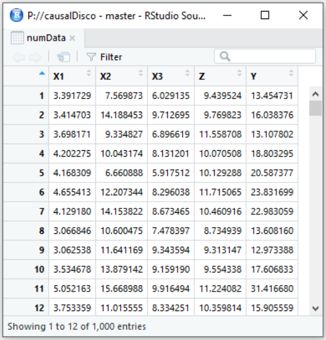
Introduction to causal discovery: CPDAGs and the PC algorithm

Anne Helby Petersen

A statistician's dream



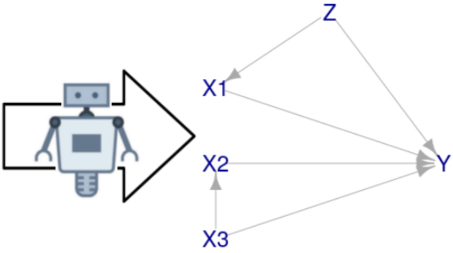
P://causalDisco - master - RStudio Sou...

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

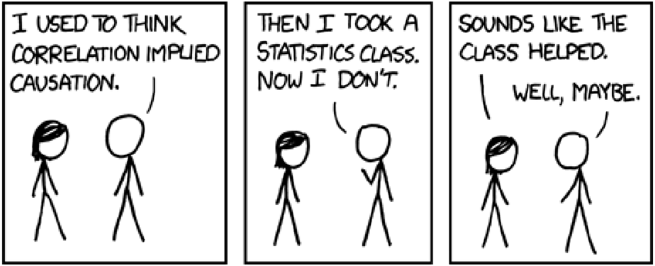
Showing 1 to 12 of 1,000 entries



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

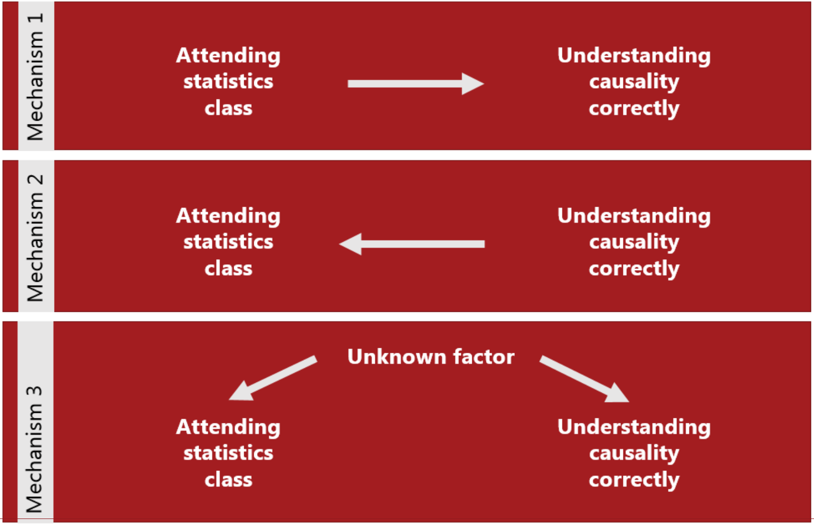
Correlation does **not** imply causation



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:



Conditional independence, association, correlation

Notation:

$X \perp\!\!\!\perp Y$ means that X is independent of Y.

$X \perp\!\!\!\perp Y \mid Z$ means that X is independent of Y conditional on Z.

Example: If the variables are jointly normally distributed, $X \perp\!\!\!\perp Y \mid Z$ means that if we fit the linear regression model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

we find $\hat{\beta}_1 \simeq 0$. In this special case, conditional independence = no association = zero correlation.

Conditional independence, association, correlation

Notation:

$X \perp\!\!\!\perp Y$ means that X is independent of Y.

$X \perp\!\!\!\perp Y \mid Z$ means that X is independent of Y conditional on Z.

Example: If the variables are jointly normally distributed, $X \perp\!\!\!\perp Y \mid Z$ means that if we fit the linear regression model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

we find $\hat{\beta}_1 \simeq 0$. In this special case, conditional independence = no association = zero correlation.

$\Rightarrow$ **(Conditional) independence is something we may be able to test from data.**

From causal structure to conditional independence

Three different minimal cases for (conditional) independence:

("d-separation" provides full definition via Markov assumption)

Case 1: Direct causal effect

$$X \rightarrow Y$$

implies that $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$ no matter what Zs are chosen.

Case 2: Causal path or confounding

$$X \rightarrow Z \rightarrow Y \quad \text{or} \quad X \leftarrow Z \leftarrow Y \quad \text{or} \quad X \leftarrow Z \rightarrow Y$$

all imply that $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$.

Case 3: Collider

$$X \rightarrow Z \leftarrow Y$$

implies that $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$. If X and Y are not adjacent, we call $X \rightarrow Z \leftarrow Y$ a *v-structure*.

From causal structure to conditional independence

Three different minimal cases for (conditional) independence:

("d-separation" provides full definition via Markov assumption)

Case 1: Direct causal effect

$$X \rightarrow Y$$

implies that $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$ no matter what Zs are chosen.

Case 2: Causal path or confounding

$$X \rightarrow Z \rightarrow Y \quad \text{or} \quad X \leftarrow Z \leftarrow Y \quad \text{or} \quad X \leftarrow Z \rightarrow Y$$

all imply that $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$.

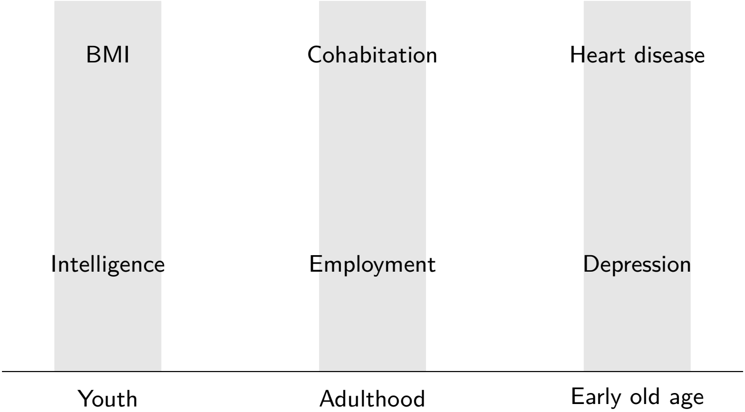
Case 3: Collider

$$X \rightarrow Z \leftarrow Y$$

implies that $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$. If X and Y are not adjacent, we call $X \rightarrow Z \leftarrow Y$ a *v-structure*.

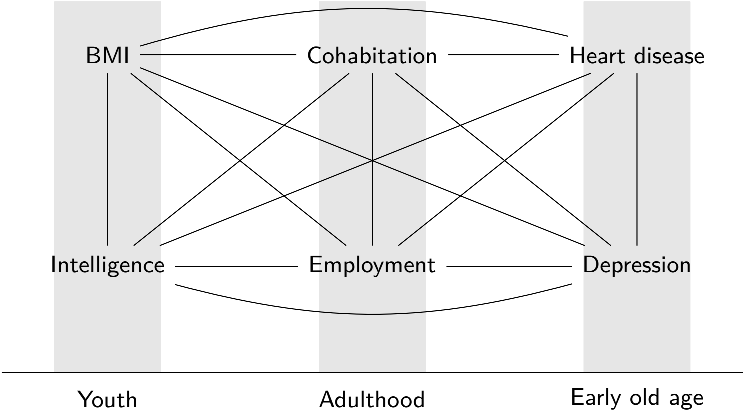
Hence can learn from data: Whether any two variables are adjacent (using 1), and if not, whether we are in case 2) or 3).

Example: Recovering a family of DAGs



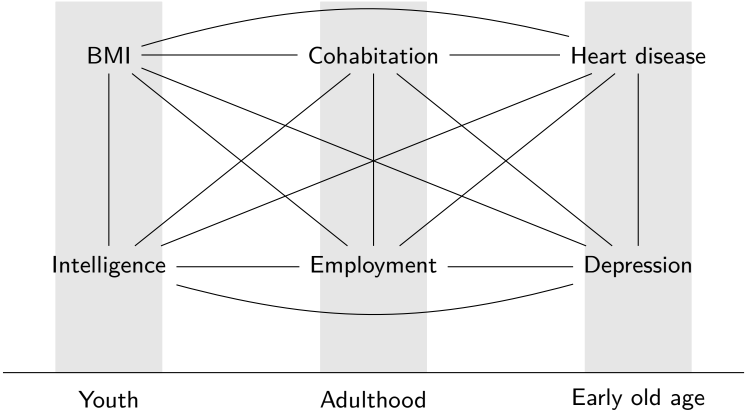
Example: Recovering a family of DAGs

First: Assume nothing, everything may be causally linked



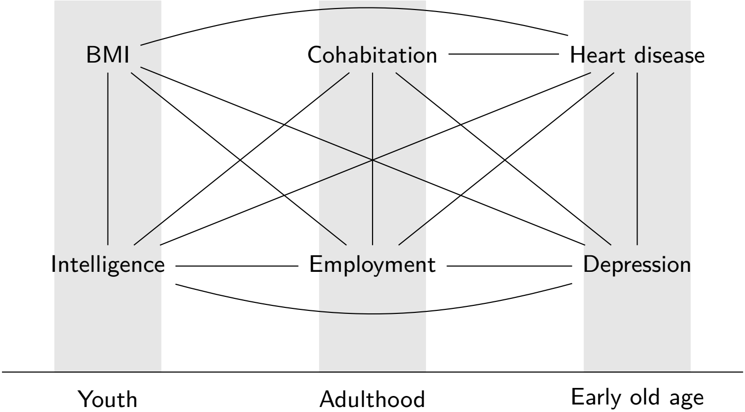
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Cohabitation.



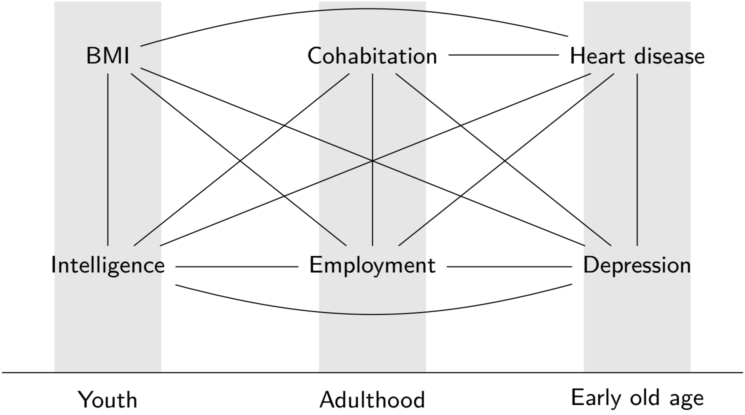
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Cohabitation.



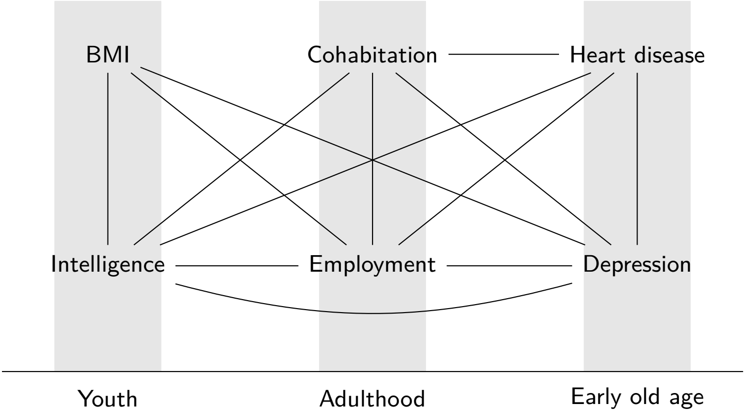
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Heart disease.



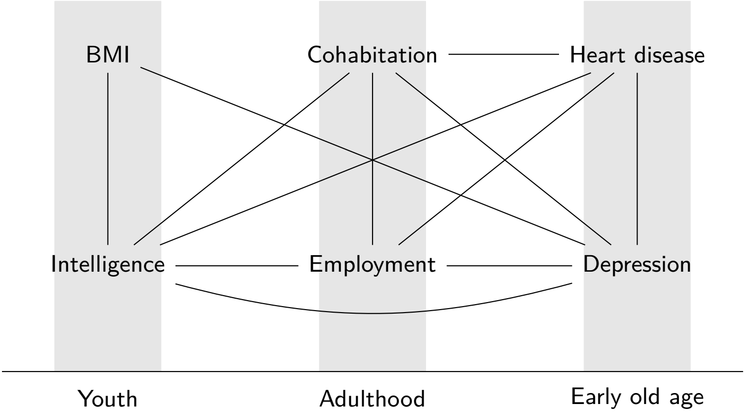
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Heart disease.



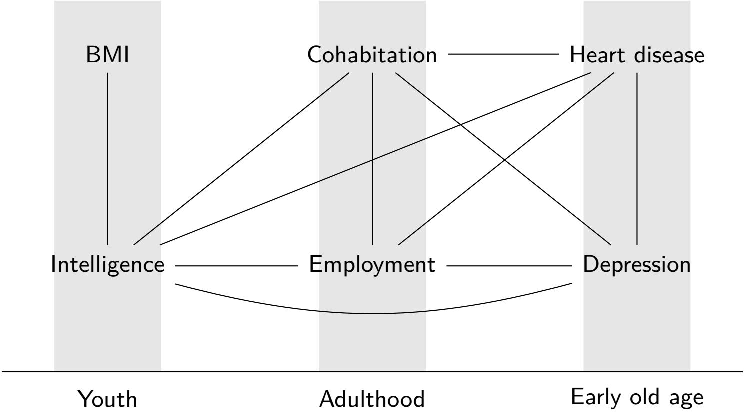
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Employment | Intelligence.



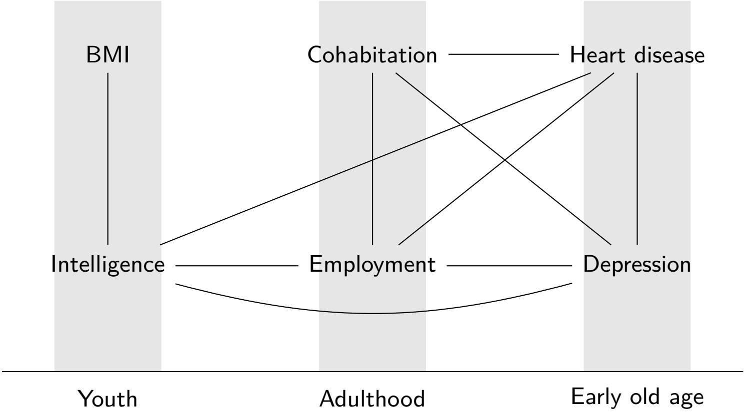
Example: Recovering a family of DAGs

Find in data: BMI $\perp\!\!\!\perp$ Depression | (Employment, Heart disease).



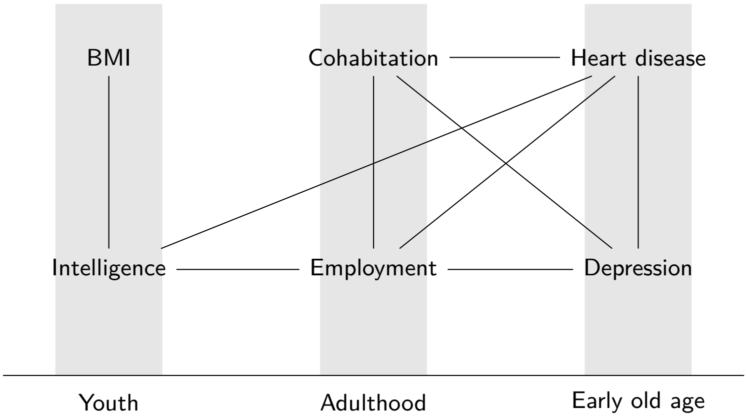
Example: Recovering a family of DAGs

Find in data: Intelligence $\perp\!\!\!\perp$ Cohabitation.



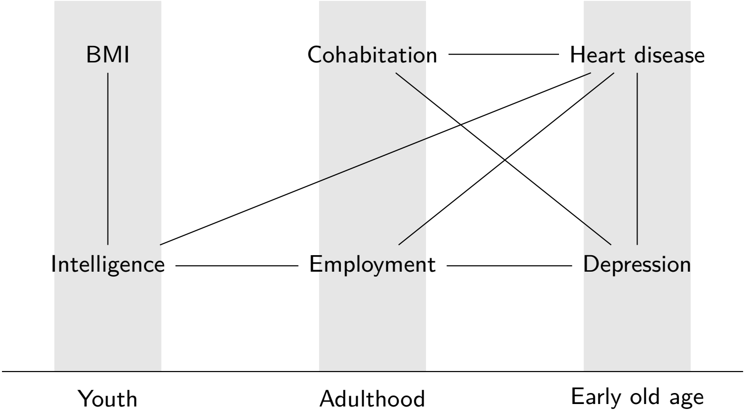
Example: Recovering a family of DAGs

Find in data: Intel. $\perp\!\!\!\perp$ Depr. | (Employm., Heart disease).



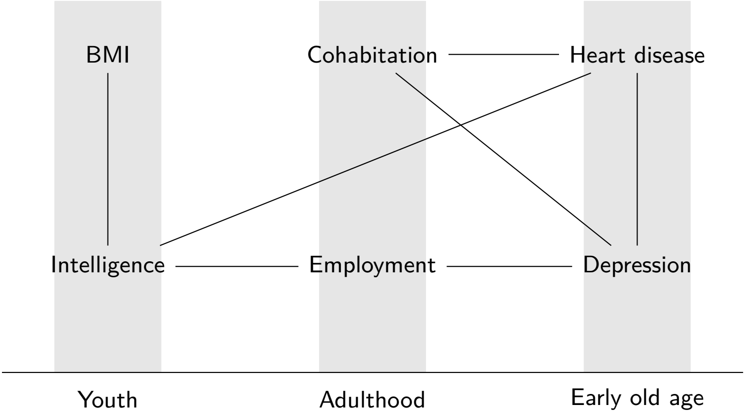
Example: Recovering a family of DAGs

Find in data: Employment $\perp\!\!\!\perp$ Cohabitation.



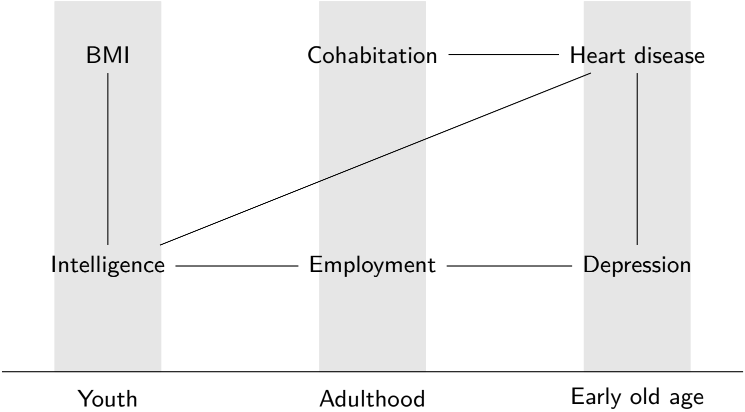
Example: Recovering a family of DAGs

Find in data: Employment $\perp\!\!\!\perp$ Heart disease.



Example: Recovering a family of DAGs

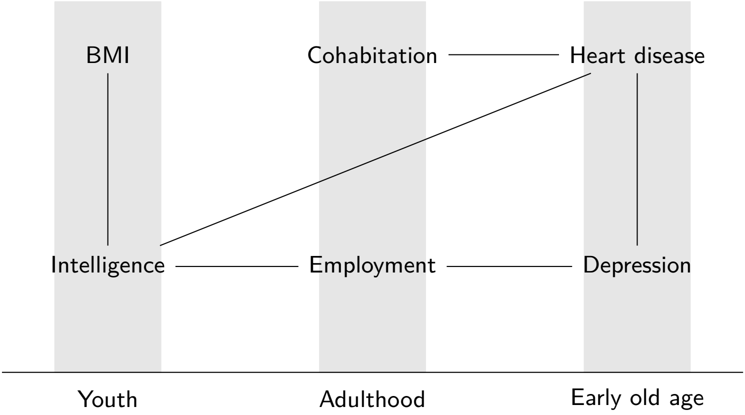
Find in data: Cohabitation $\perp\!\!\!\perp$ Depression | Heart disease.



Example: Recovering a family of DAGs

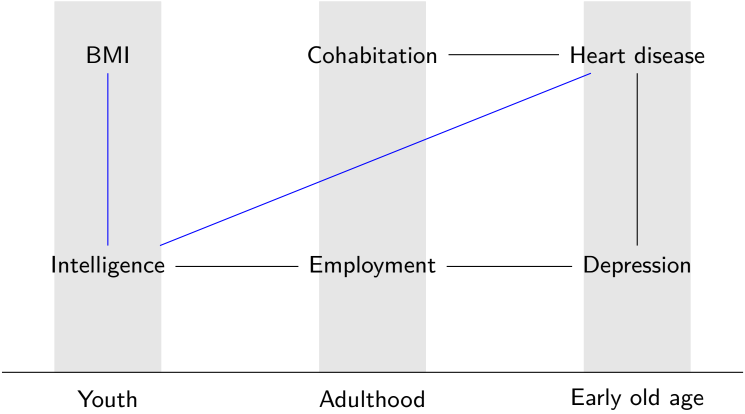
Find in data: No other (conditional) independencies!

Next step: Look for v-structures.



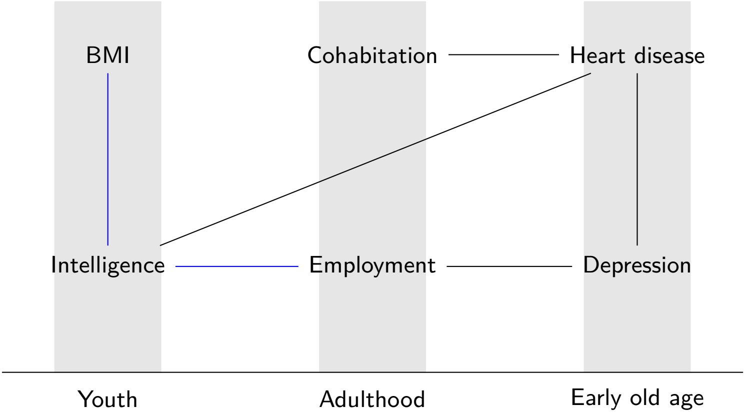
Example: Recovering a family of DAGs

Potential v-structure: BMI – Intel. – Heart disease.
Find in data: BMI $\perp\!\!\!\perp$ Heart disease | Intel. **Not v-struct.**



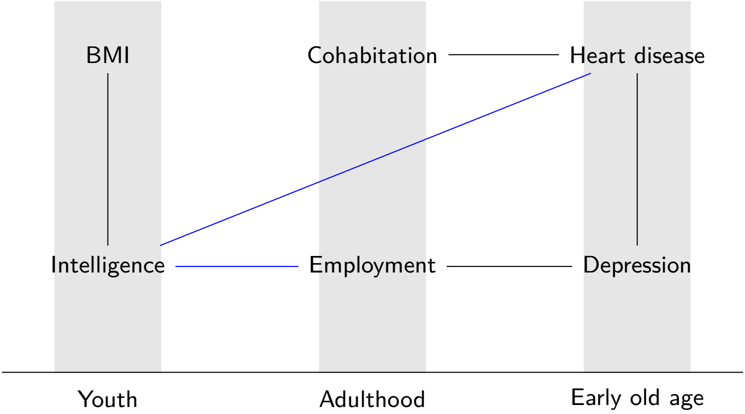
Example: Recovering a family of DAGs

Potential v-structure: BMI – Intel. – Employm.
Find in data: BMI $\perp\!\!\!\perp$ Employm. | Intel. **Not v-struct.**



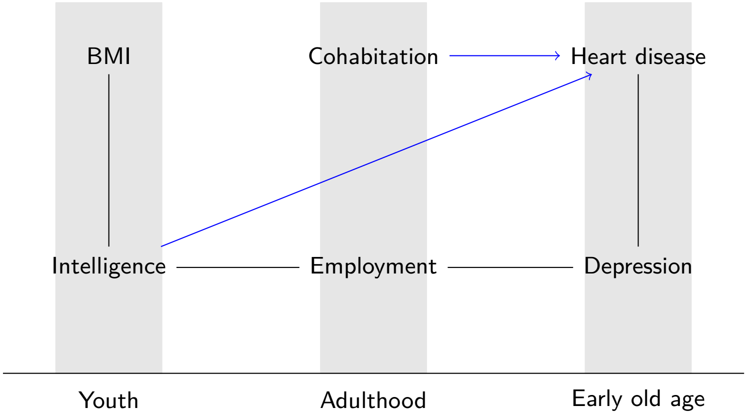
Example: Recovering a family of DAGs

Potential v-structure: Heart disease – Intel. – Employm.
Find in data: Heart disease $\perp\!\!\!\perp$ Employm. | Intel. **Not v-struct.**



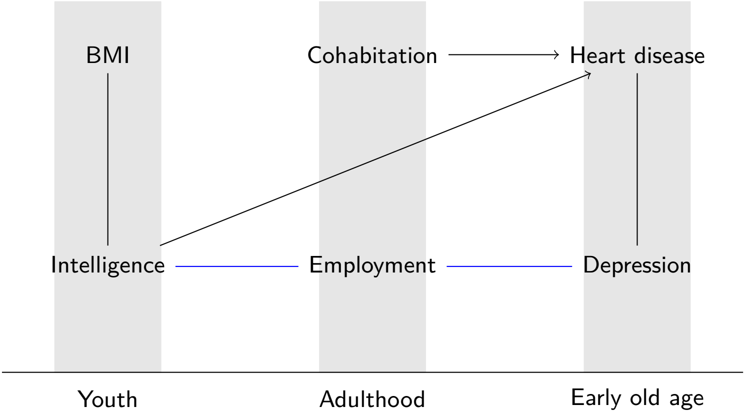
Example: Recovering a family of DAGs

Potential v-structure: Cohab. – Hearth disease – Intel.
Find in data: Cohab. $\not\perp\!\!\!\perp$ Intel. | Hearth disease. **V-struct.!**



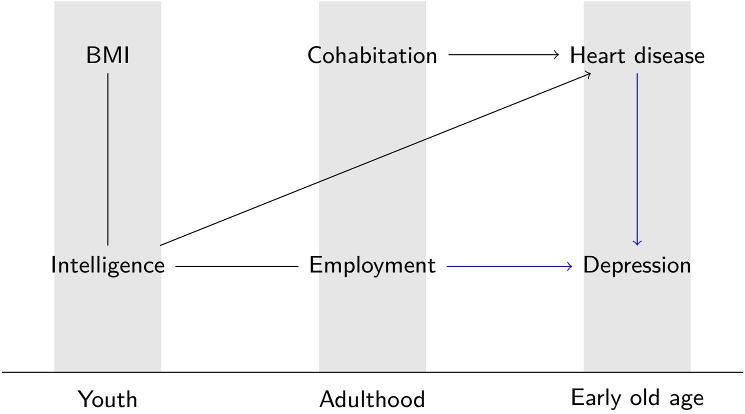
Example: Recovering a family of DAGs

Potential v-structure: Intel. − Employm. − Dep.
Find in data: Intel. $\perp\!\!\!\perp$ Dep. | Employm. **Not v-struct.**



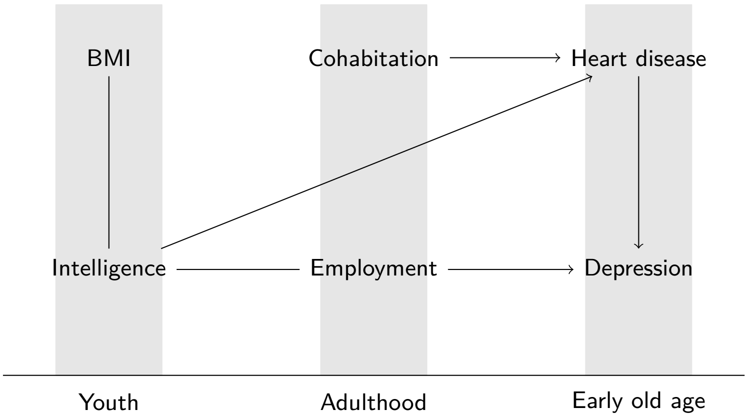
Example: Recovering a family of DAGs

Potential v-structure: Employm. – Dep. – Heart disease
Find in data: Employm. $\perp\!\!\!\perp$ Heart disease | Dep. **V-struct.!**



Example: Recovering a family of DAGs

No more potential v-structures to consider.
Final graph! Note: Some edges still unoriented.



The PC algorithm (Spirtes & Glymour 1991)

Peter-Clark (PC) algorithm summary

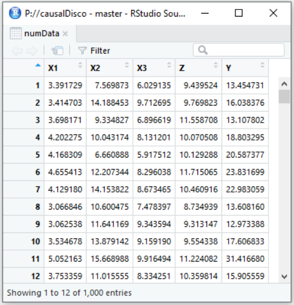
Input: Information about conditional independencies[a]

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 such that $A \perp\!\!\!\perp B \mid S$. If such an S exists: Remove edge between A and B.
3. Look for v-structures: For each triple $A - B - C$ with A and C non-adjacent, if B is not in any separating set for (A, C), orient as $A \rightarrow B \leftarrow C$.
4. Apply additional orientation rules avoiding introducing new v-structures and cycles.

Output: Family ("equivalence class") of DAGs - completed partially directed acyclic graph (CPDAG).

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

A statistician's dream version 2.0



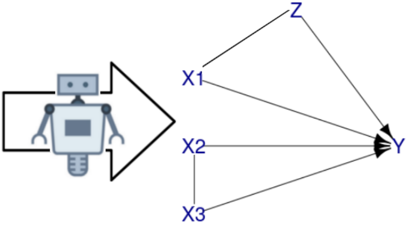
P://causalDisco - master - RStudio Sou...

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

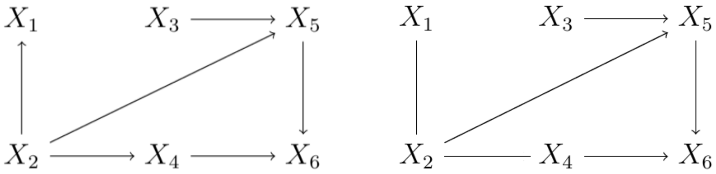


Goal: Estimate **CPDAG** by analyzing data (i.e., causal discovery).

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/CPDAG).

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

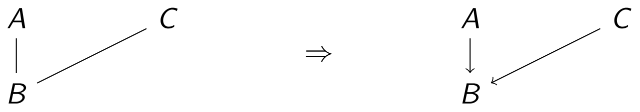
CPDAG 101: A family of DAGs



- We can divide all possible DAGs into **families** (equivalence classes) according to what conditional independencies they entail.
- All DAGs in a family will have the **same adjacencies** + **same v-structures**, and they can be represented by a (unique) completed partially directed acyclic graph (**CPDAG**).
- CPDAGs can have both **directed** ($\rightarrow$) and **undirected** ($-$) edges.
- **CPDAG interpretation:** Directed edges are interpreted as for DAGs. Undirected edges mean that some family members orient it one way, others the other way.

PC orientation rules

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

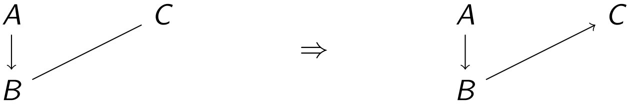


Next, recursively apply **three additional rules** (next slide) until no further changes are made.

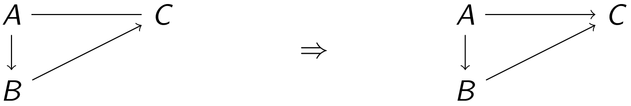
These rules are **sound and complete** (in the large sample limit): No incorrect orientations occur, and no further orientations can be made (Meek 1995).

Meek's orientation rules

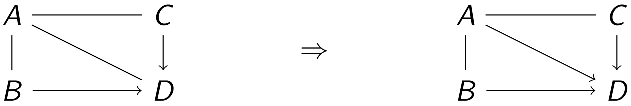
R1: Avoid introducing new v-structures (directly):



R2: Avoid introducing cycles.



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



PC algorithm assumptions and guarantees

Assumptions:

- Acyclic data generating mechanisms: No feedback loops
- No unobserved confounding
- Faithfulness: Cond. independencies in data $\Rightarrow$ DAG structure
- No conditioning on unobserved colliders

Under these assumptions + a valid test of conditional independence, PC is mathematically proven to be

correct: no false claims, *and*

complete: no further causal structure can be inferred without additional assumptions

PC algorithm assumptions and guarantees

Assumptions:

- Acyclic data generating mechanisms: No feedback loops
- No unobserved confounding
- Faithfulness: Cond. independencies in data $\Rightarrow$ DAG structure
- No conditioning on unobserved colliders

Under these assumptions + a valid test of conditional independence, PC is mathematically proven to be

correct: no false claims, *and*

complete: no further causal structure can be inferred without additional assumptions

... but in practice: **Assumptions may be violated, conditional independence tests subject to statistical error.**

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:

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

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

and $\text{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta_1 = 0$ in the linear regression model

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + \epsilon_i$$

[1] Up to statistical uncertainty...

Conditional independence testing

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

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

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

and $\mathrm{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta_1 = 0$ in the linear regression model

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + \epsilon_i$$

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:

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

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

and $\text{cor}(X, Y \mid Z) = 0$ is equivalent with testing $H_0 : \beta_1 = 0$ in the linear regression model

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + \epsilon_i$$

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*.