Causal learning

Marloes Maathuis
ETH Zurich

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by sex

References: Yule (1903), Simpson (1951), Hernan, Clayton and Keiding (2011), Pearl (2014).

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by sex

- Among males, treatment is better

References: Yule (1903), Simpson (1951), Hernan, Clayton and Keiding (2011), Pearl (2014)

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by sex

- Among males, treatment is better
- Among females, treatment is better

References: Yule (1903), Simpson (1951), Hernan, Clayton and Keiding (2011), Pearl (2014)

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by sex

- Among males, treatment is better
- Among females, treatment is better
- Overall, placebo is better

References: Yule (1903), Simpson (1951), Hernan, Clayton and Keiding (2011), Pearl (2014)

Simpson's paradox

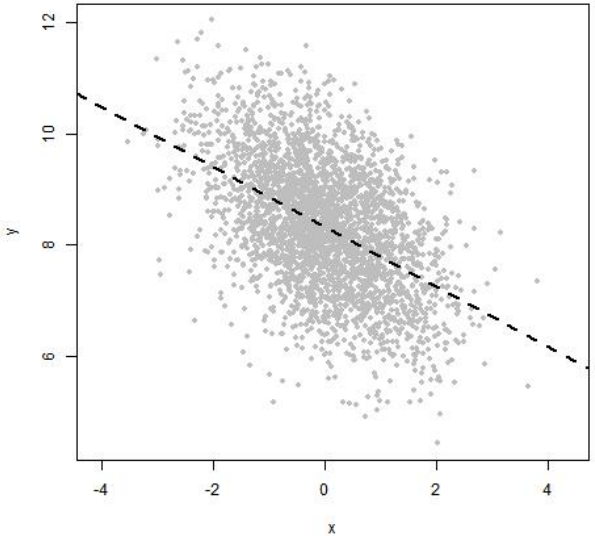
	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by sex

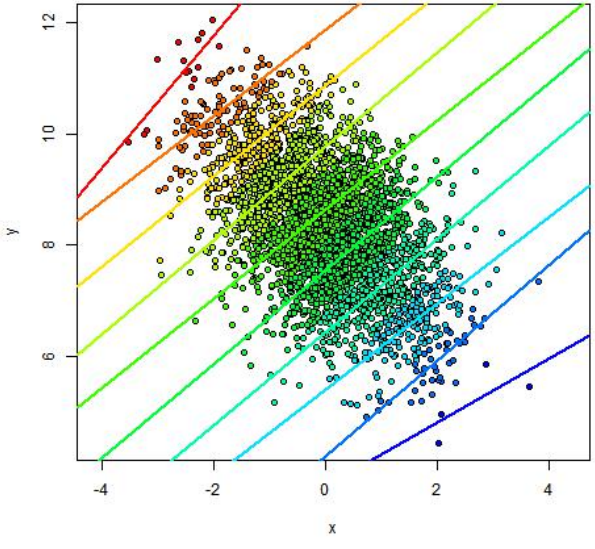
- Among males, treatment is better
- Among females, treatment is better
- Overall, placebo is better

References: Yule (1903), Simpson (1951), Hernan, Clayton and Keiding (2011), Pearl (2014)

Simpson's paradox for continuous variables



Color encodes third variable Z



Source: http://www.r-bloggers.com/fun-with-simpsons-paradox-simulating-confounders/

What should we conclude?

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

What should we conclude?

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Adjust for sex

Treatment works for both males and females -> it works

What should we conclude?

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Adjust for sex

Treatment works for both males and females -> it works

Replace sex by blood pressure (BP)

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

What should we conclude?

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Adjust for sex

Treatment works for both males and females -> it works

Replace sex by blood pressure (BP)

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

Do not adjust for BP

Treatment does not work

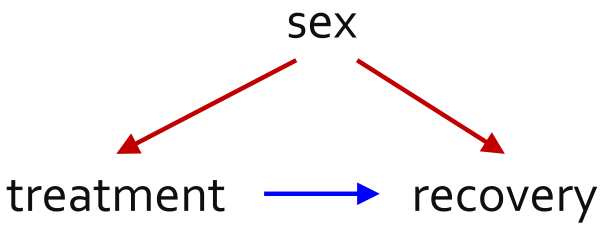
Simpson's paradox and causal diagrams

- Same numbers, but different conclusions?
  - Conclusions must use additional information: story behind the data, causal assumptions

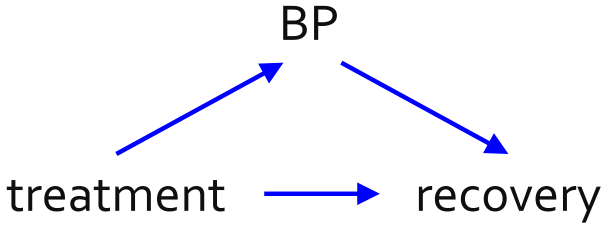
Simpson's paradox and causal diagrams

- Same numbers, but different conclusions?
  - Conclusions must use additional information: story behind the data, causal assumptions

- We want to know the causal effect of treatment on recovery. Possible scenarios:



sex is a confounder
→ adjust



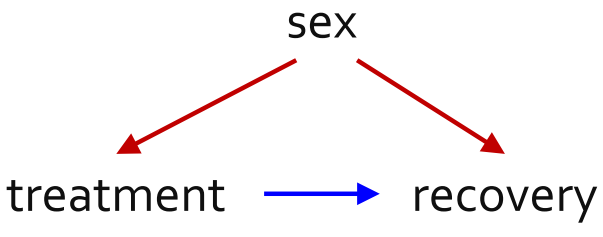
BP is an intermediate variable
→ do not adjust

Or……

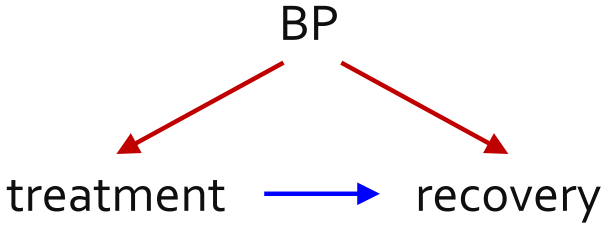
Simpson's paradox and causal diagrams

- Same numbers, but different conclusions?
  - Conclusions must use additional information: story behind the data, causal assumptions

- We want to know the causal effect of treatment on recovery. Possible scenarios:



sex is a confounder
→ adjust



BP is an confounder
→ adjust

Take home message

- Simpson's paradox:
  - The interpretation of a parameter in a model depends on the other variables in the model
  - Simpson's paradox is an extreme case where the sign flips

Take home message

- Simpson's paradox:
  - The interpretation of a parameter in a model depends on the other variables in the model
  - Simpson's paradox is an extreme case where the sign flips

- Causality:
  - Answering causal questions from observational data requires causal assumptions
  - Such assumptions can be formalized in a causal graph

Outline of this talk

1. Terminology
2. Identification and estimation of causal effects when the causal graph is known, using adjustment
3. What can we do if the causal graph is unknown?

Causal versus non-causal questions

- Non-causal questions are about predictions in the same system:
  - Predict the cancer rate among smokers
  - Finding biomarkers for a disease
  - Classification of images
  - ...
- Causal questions are about the mechanism behind the data or about predictions after an intervention to the system:
  - Does smoking cause lung cancer?
  - Finding therapeutic targets for a disease
  - Predicting the growth of the Corona epidemic after imposing new regulations
  - ...

Experimental versus observational data

- Causal questions are ideally answered by randomized controlled experiments. Examples:
  - agricultural experiments
  - clinical trials to test new drugs
  - A/B testing

  Randomization ensure there is no confounding
- Sometimes such experiments are impossible, as they may be:
  - unethical (smoking)
  - infeasible (global warming)
  - expensive / time consuming (gene knock-outs)

Goal: estimate causal effects from observational data

Interventional notion of total causal effect

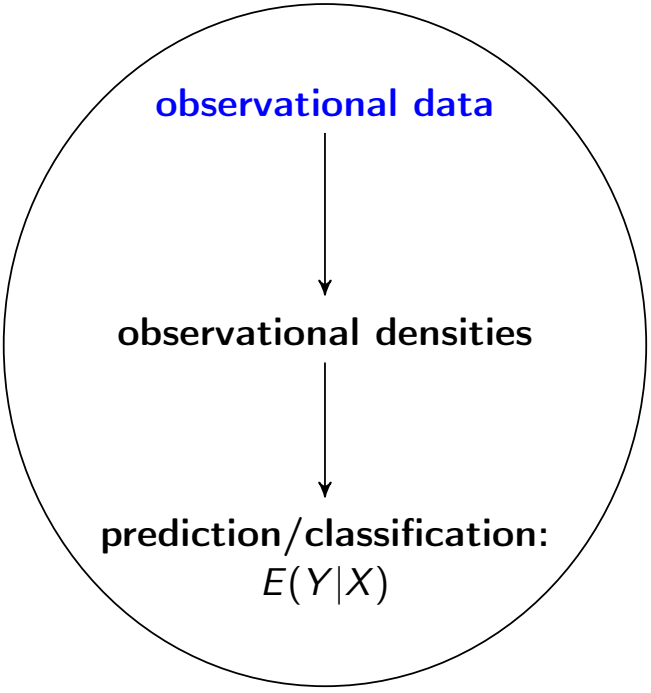
- There is a causal effect of X on Y if there exist $x \neq x'$ such that $f(y|do(X = x)) \neq f(y|do(X = x'))$
  - $do(X = x)$ or $do(x)$ denotes setting X to the value x by an outside intervention, uniformly over the entire population
  - $f(y|do(x))$ is the post-intervention density of y after $do(x)$
- The total causal effect of X on Y can be summarized as, e.g., $\frac{\partial}{\partial x} E(Y|do(x))$

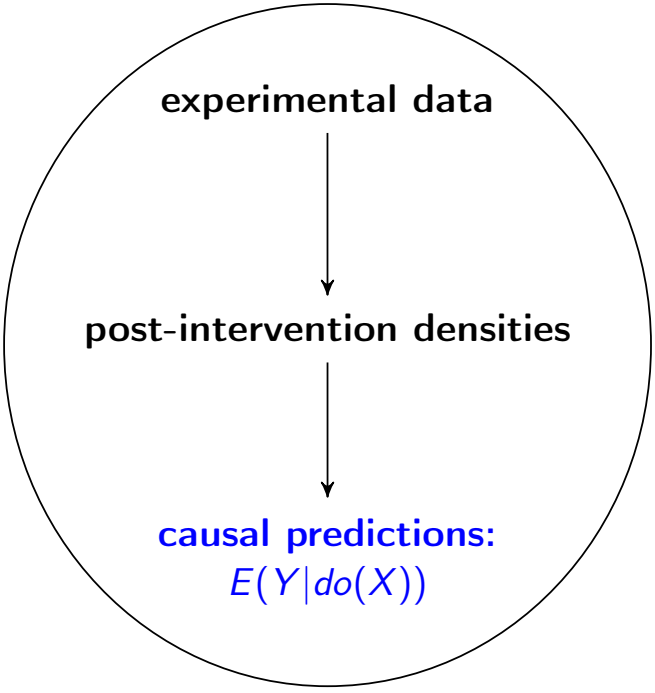
Example: effect of smoking on lung cancer

- $Y = 1$ if a person has lung cancer; $Y = 0$ otherwise
- $X = 1$ if a person smokes; $X = 0$ otherwise
- $E(Y|do(X = 1))$ is the post-intervention lung cancer rate if everybody were forced to smoke
- Total causal effect of X on Y:

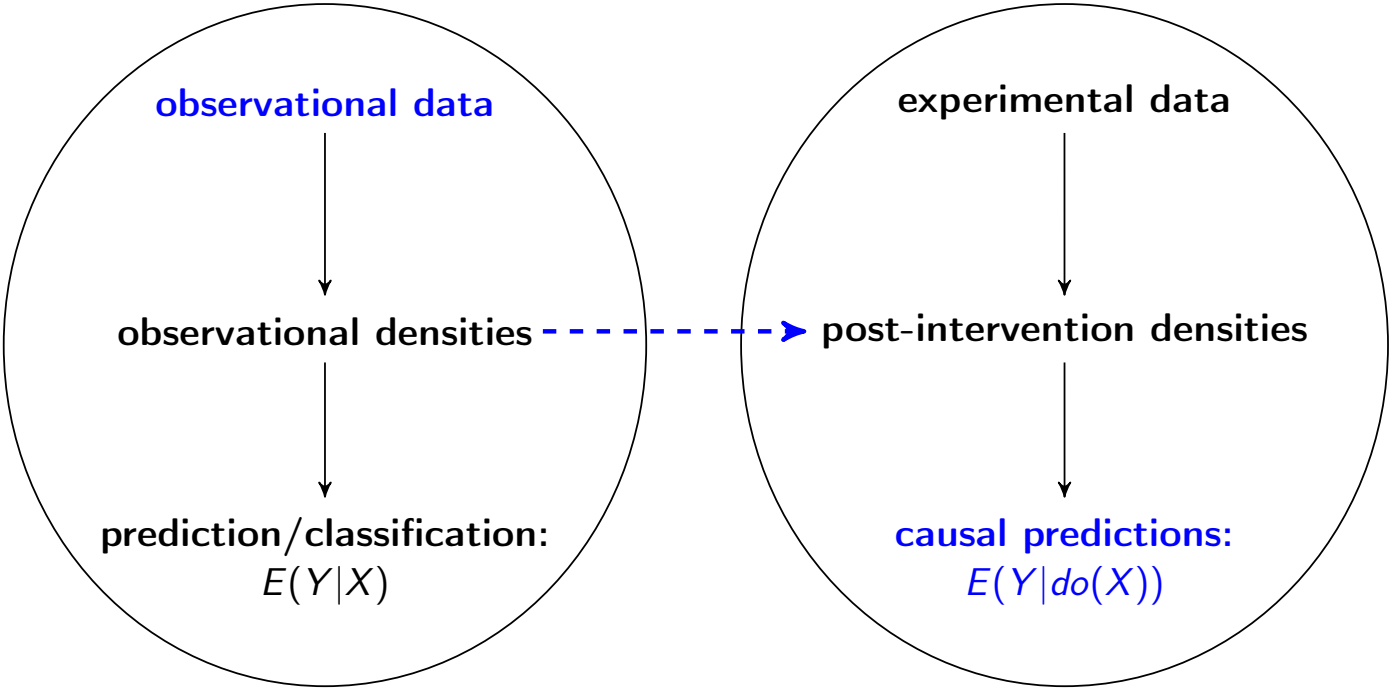
$$E(Y|do(X = 1)) - E(Y|do(X = 0))$$

How can we estimate causal effects from observational data?





How can we estimate causal effects from observational data?



Common assumption:
Data come from a known causal directed acyclic graph (DAG)

DAGs

- DAG: every vertex represents a variable, there are only directed edges and no directed cycles.
- A density f is compatible with a DAG $\mathcal{D}$ if it factorizes wrt $\mathcal{D}$:

$$f(\mathbf{v}) = \prod_{V \in \mathbf{V}} f(v|pa(v, \mathcal{D}))$$

- Example: every density $f(x, y)$ is compatible with the DAGs

$$X \to Y \qquad \text{since} \qquad f(x, y) = f(x)f(y|x)$$

and

$$X \leftarrow Y \qquad \text{since} \qquad f(x, y) = f(y)f(x|y)$$

Causal DAGs

- A density f is compatible with a causal DAG $\mathcal{D}$ if for any $\mathbf{X} \subseteq \mathbf{V}$ it satisfies the truncated factorization formula:

$$f(\mathbf{v} \setminus \mathbf{x} \,|\, do(\mathbf{X} = \mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{X}} f(v | pa(v, \mathcal{D})) \Big|_{\mathbf{X}=\mathbf{a}}$$

Causal DAGs

- A density f is compatible with a causal DAG $\mathcal{D}$ if for any $\mathbf{X} \subseteq \mathbf{V}$ it satisfies the truncated factorization formula:

$$f(\mathbf{v} \setminus \mathbf{x} \,|\, do(\mathbf{X} = \mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{X}} f(v | pa(v, \mathcal{D})) \Big|_{\mathbf{X}=\mathbf{a}}$$

post-intervention density

conditional densities

Causal DAGs

- A density f is compatible with a causal DAG $\mathcal{D}$ if for any $\mathbf{X} \subseteq \mathbf{V}$ it satisfies the truncated factorization formula:

$$f(\mathbf{v}\backslash\mathbf{x} \,|\, do(\mathbf{X} = \mathbf{a})) = \prod_{V \in \mathbf{V}\backslash\mathbf{X}} f(v|pa(v, \mathcal{D}))\Big|_{\mathbf{X}=\mathbf{a}}$$

- Key assumption is autonomy/invariance:
the conditional distribution of a node given its parents is invariant to interventions at other nodes

Causal DAGs

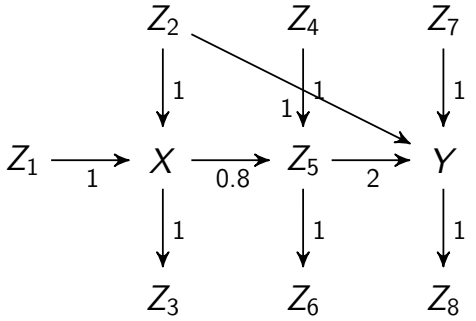
- A density f is compatible with a causal DAG $\mathcal{D}$ if for any $\mathbf{X} \subseteq \mathbf{V}$ it satisfies the truncated factorization formula:

$$f(\mathbf{v} \backslash \mathbf{x} \,|\, do(\mathbf{X} = \mathbf{a})) = \prod_{V \in \mathbf{V} \backslash \mathbf{X}} f(v|pa(v, \mathcal{D}))\Big|_{\mathbf{X}=\mathbf{a}}$$

- Key assumption is autonomy/invariance:
the conditional distribution of a node given its parents is invariant to interventions at other nodes
- As causal DAGs, $X \to Y$ and $X \leftarrow Y$ are markedly different:

$$\begin{aligned} X \to Y: &\quad f(y|do(x=1)) = f(y|x=1) \\ X \leftarrow Y: &\quad f(y|do(x=1)) = f(y) \end{aligned}$$

Example: linear structural equation model

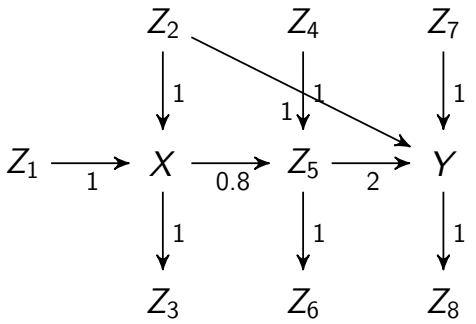


Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- 0.8*X + Z4 + rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

The resulting distribution is compatible with the causal DAG

Example: linear structural equation model

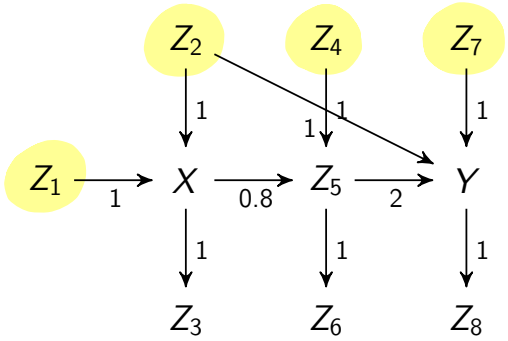


Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- 0.8*X + Z4 + rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

The resulting distribution is compatible with the causal DAG

Example: linear structural equation model

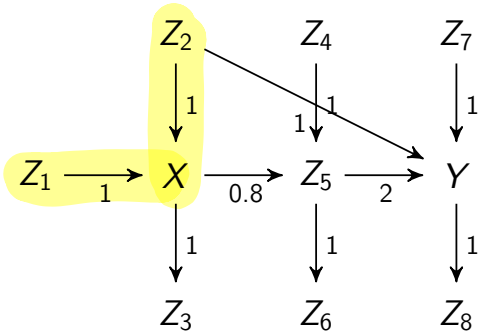


Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- 0.8*X + Z4 + rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

The resulting distribution is compatible with the causal DAG

Example: linear structural equation model

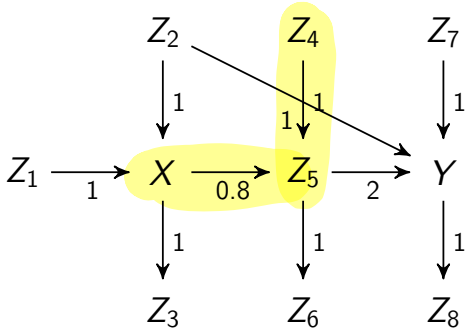


Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- 0.8*X + Z4 + rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

The resulting distribution is compatible with the causal DAG

Example: linear structural equation model

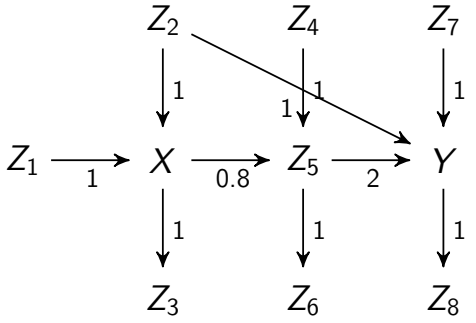


Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- 0.8*X + Z4 + rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

The resulting distribution is compatible with the causal DAG

Example: $do(Z_5 = 3)$



Each variable is generated as a linear function of its parents:

```
n  <- 100000                          # sample size
Z1 <- rnorm(n);  Z2 <- rnorm(n)
Z4 <- rnorm(n);  Z7 <- rnorm(n)
X  <- Z1 + Z2 + rnorm(n)
Z5 <- rep(3,n)                        # before: 0.8*X+Z4+rnorm(n)
Y  <- Z2 + 2*Z5 + Z7 + rnorm(n)
Z3 <- X  + rnorm(n)
Z6 <- Z5 + rnorm(n)
Z8 <- Y  + rnorm(n)
```

Identification of causal effects

- Identification: Given a density f that is compatible to a causal DAG, can we write $f(\mathbf{y}|do(\mathbf{x}))$ as a function of conditional densities that we can estimate?
- Methods: truncated factorization formula, back-door/adjustment formula, front-door formula, ID algorithm (Pearl, Tian, Shpitser, ...)
- We focus on identifiability via adjustment

Adjustment

- Definition: $\mathbf{S}$ is a valid adjustment set for $(\mathbf{X}, \mathbf{Y})$ in a causal DAG $\mathcal{D}$ if for any f compatible with $\mathcal{D}$:

$$f(\mathbf{y}|do(\mathbf{x})) = \int_{\mathbf{S}} f(\mathbf{y}|\mathbf{x}, \mathbf{s}) f(\mathbf{s}) d\mathbf{s}$$

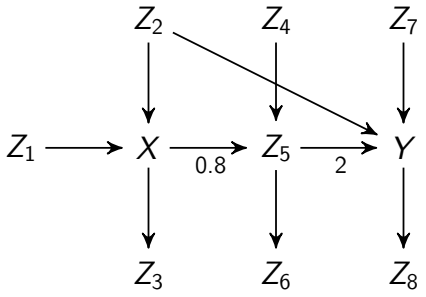
Adjustment

- Definition: $\mathbf{S}$ is a valid adjustment set for $(\mathbf{X}, \mathbf{Y})$ in a causal DAG $\mathcal{D}$ if for any f compatible with $\mathcal{D}$:

$$f(\mathbf{y}|do(\mathbf{x})) = \int_{\mathbf{S}} f(\mathbf{y}|\mathbf{x}, \mathbf{s}) f(\mathbf{s}) d\mathbf{s}$$

- For singleton X and Y:
  - If $Y \notin pa(X, \mathcal{D})$, then $pa(X, \mathcal{D})$ is a valid adjustment set
  - In a linear system, the total effect $\frac{\partial}{\partial x} E(Y|do(x))$ is the coefficient of X in the linear regression $Y \sim X + \mathbf{S}$

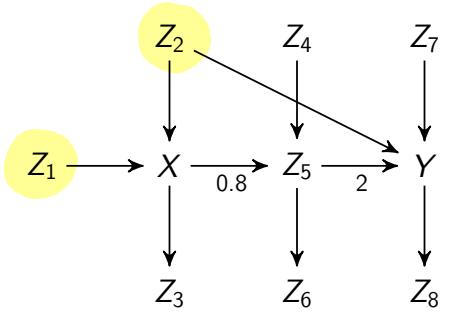
Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

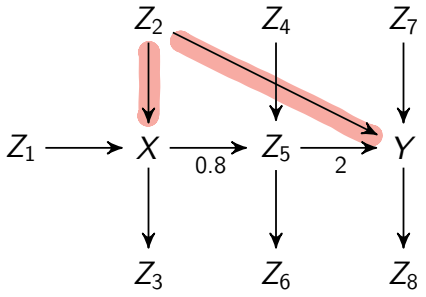
Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

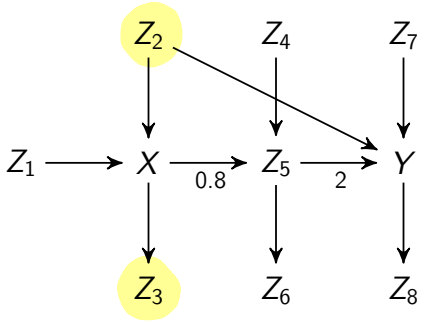
Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

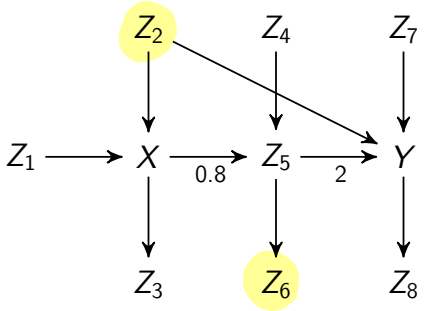
Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

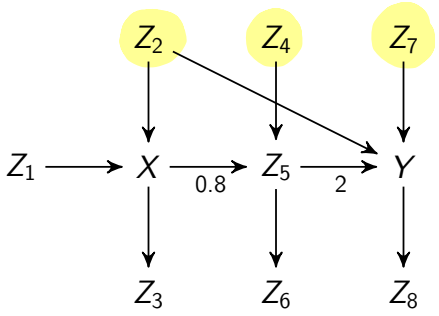
Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

Can we estimate the total effect by adjusted regression?



True total causal effect of X on Y is $0.8 \cdot 2 = 1.6$

```
> lm(Y~X+Z1+Z2)$coeff[2]
1.611                          # S = pa(X), OK
> lm(Y~X)$coeff[2]
1.940                          # S = empty, invalid
> lm(Y~X+Z2+Z3)$coeff[2]
1.606                          # S = {Z2,Z3}, OK
> lm(Y~X+Z2+Z6)$coeff[2]
0.542                          # S = {Z2,Z6}, invalid
> lm(Y~X+Z2+Z4+Z7)$coeff[2]
1.604                          # S = {Z2,Z4,Z7}, OK
```

What are valid adjustment sets?

Common ideas about adjustment:

- adjusting for more variables is better
- one should adjust for all variables related to both X and Y
- adjusting for pre-treatment variables is always safe
- adjusting for descendants of X is always bad
- ...

What are valid adjustment sets?

Common ideas about adjustment:

- adjusting for more variables is better
- one should adjust for all variables related to both X and Y
- adjusting for pre-treatment variables is always safe
- adjusting for descendants of X is always bad
- ...

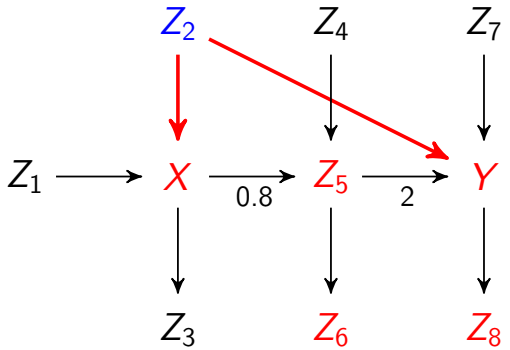
These are generally false.
What to do instead? Use graphical criteria!
Backdoor criterion (Pearl) and adjustment criterion

Adjustment criterion for DAGs

Theorem (Shpitser et al '10, Perković et al '18):
$\mathbf{Z}$ is a valid adjustment set for $(\mathbf{X}, \mathbf{Y})$ in a causal DAG $\mathcal{D}$ iff the following two conditions hold:

- $\mathbf{Z} \cap \mathit{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \emptyset$
- $\mathbf{Z}$ blocks all proper non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}$

Example:



Valid adjustment sets are $\{Z_2\} \cup \mathbf{S}$ for $\mathbf{S} \subseteq \{Z_1, Z_3, Z_4, Z_7\}$

What about efficiency?

Among all valid adjustment sets, which set provides the optimal asymptotic variance for the causal effect estimate?

Define O-set $:= pa(cn(\mathbf{X}, \mathbf{Y}, \mathcal{D})) \setminus forb(\mathbf{X}, \mathbf{Y}, \mathcal{D})$

Theorem (Henckel et al '19): Let $\mathbf{Y} \subseteq de(\mathbf{X}, \mathcal{D})$. Then

- The O-set is a valid adjustment set wrt $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ iff there exists a valid adjustment set
- The O-set is asymptotically optimal for causal linear models

What about efficiency?

Among all valid adjustment sets, which set provides the optimal asymptotic variance for the causal effect estimate?

Define O-set $:= pa(cn(\mathbf{X}, \mathbf{Y}, \mathcal{D})) \setminus forb(\mathbf{X}, \mathbf{Y}, \mathcal{D})$

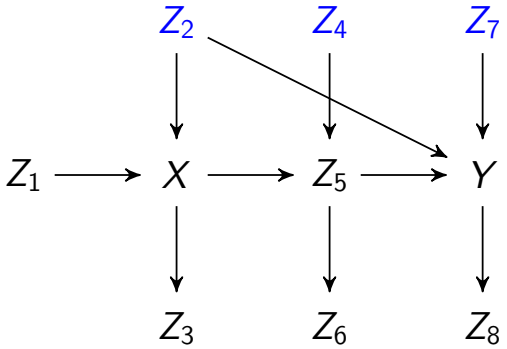
Theorem (Henckel et al '19): Let $\mathbf{Y} \subseteq de(\mathbf{X}, \mathcal{D})$. Then

- The O-set is a valid adjustment set wrt $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ iff there exists a valid adjustment set
- The O-set is asymptotically optimal for causal linear models

The latter result was generalized to non-parametrically adjusted estimators of interventional means (Rotnitzky & Smucler '20)

The O-set

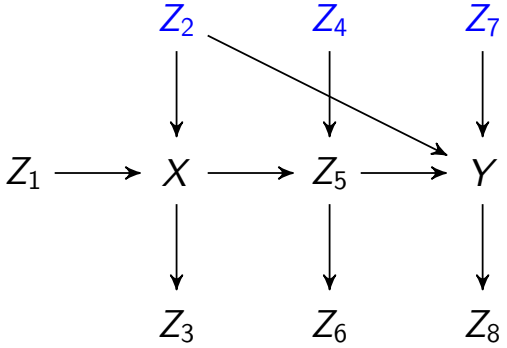
- Example:



$$\mathbf{O}(X, Y, \mathcal{D}) = \{Z_2, Z_4, Z_7\}$$

The **O**-set

- Example:

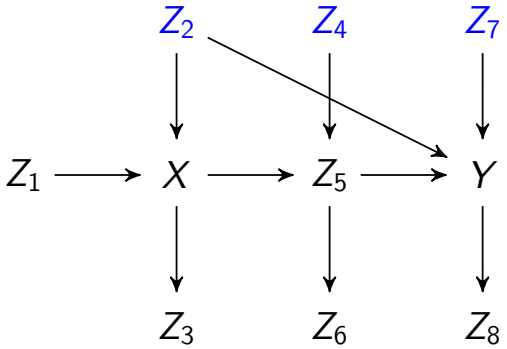


$$\mathbf{O}(X, Y, \mathcal{D}) = \{Z_2, Z_4, Z_7\}$$

- Intuition: regression $Y \sim X + \mathbf{S}$:
  S should explain a lot of variance of Y
  S should have small correlation with X

The **O**-set

- Example:



$$\mathbf{O}(X, Y, \mathcal{D}) = \{Z_2, Z_4, Z_7\}$$

- Intuition: regression $Y \sim X + \mathbf{S}$:
  S should explain a lot of variance of Y
  S should have small correlation with X
- In particular, $pa(X, \mathcal{D})$ is typically bad in terms of variance

What if the causal DAG is unknown?

Approach 1: Hypothesize possible causal DAGs

- Drawing DAGs formalizes the causal assumptions
- Each hypothesized DAG can be used to estimate causal effect of interest
- Allows sensitivity analysis and informed discussion

What if the causal DAG is unknown?

Approach 2: Learn the DAG from data

- A DAG encodes d-separations
- Several DAGs can encode the same d-separations.
  Such DAGs form a Markov equivalence class. Example:

$$
\begin{array}{lll}
X_1 \to X_2 \to X_3: & X_1 \not\perp X_3 & X_1 \perp X_3 | X_2 \\
X_1 \leftarrow X_2 \to X_3: & X_1 \not\perp X_3 & X_1 \perp X_3 | X_2 \\
X_1 \leftarrow X_2 \leftarrow X_3: & X_1 \not\perp X_3 & X_1 \perp X_3 | X_2 \\
X_1 \to X_2 \leftarrow X_3: & X_1 \perp X_3 & X_1 \not\perp X_3 | X_2
\end{array}
$$

- A Markov equivalence class can be described by a CPDAG.
  A CPDAG is identifiable from observational data.

What if the causal DAG is unknown?

Approach 2: Learn the DAG from data

- A DAG encodes d-separations
- Several DAGs can encode the same d-separations.
  Such DAGs form a Markov equivalence class. Example:

$X_1 \to X_2 \to X_3$:	$X_1 \not\perp X_3$	$X_1 \perp X_3 \mid X_2$
$X_1 \leftarrow X_2 \to X_3$:	$X_1 \not\perp X_3$	$X_1 \perp X_3 \mid X_2$
$X_1 \leftarrow X_2 \leftarrow X_3$:	$X_1 \not\perp X_3$	$X_1 \perp X_3 \mid X_2$
$X_1 \to X_2 \leftarrow X_3$:	$X_1 \perp X_3$	$X_1 \not\perp X_3 \mid X_2$

- A Markov equivalence class can be described by a CPDAG.
  A CPDAG is identifiable from observational data.

Causal structure learning without hidden variables

- Assumption: F is Markov and faithful to the causal DAG
  $\{\text{d-sep in the DAG}\} = \{\text{conditional independencies in } F\}$
- Constraint-based methods:
  - Use conditional independencies in observational distribution
  - Example: PC (Spirtes et al '00, Kalisch & Bühlmann '07, Colombo & MM '14)
- Score-based methods:
  - A score function is optimized over the space of DAGs/CPDAGs
  - Example: GES (Chickering '02, Nandy et al '18)
- Hybrid methods:
  - Examples: MMHC (Tsmardinos et al '06), NSDIST (Han et al '16), ARGES (Nandy et al '18)

What if there are hidden variables?

- Constraint-based methods that allow arbitrarily many hiddens:
  - FCI (Spirtes et al '00)
  - RFCI (Colombo et al '12)
  - FCI+ (Claassen et al '13)

  Output is a PAG

What if there are hidden variables?

- Constraint-based methods that allow arbitrarily many hiddens:
  - FCI (Spirtes et al '00)
  - RFCI (Colombo et al '12)
  - FCI+ (Claassen et al '13)

  Output is a PAG

- Impose conditions on hiddens: allow a few hiddens that affect many of the observed variables:
  - Precision matrix has low rank + sparse structure
  - LRpS-GES (Frot et al '19)

  Output is a CPDAG

Overview of graphical criteria for more general graphs

	DAG	CPDAG	PAG
Backdoor criterion Pearl '93	✓		
Adjustment criterion Shpitser et al '12, Perković et al '18	✓		
Generalized backdoor criterion MM & Colombo '15	✓	✓	✓
Generalized adjustment criterion Perković et al '15, '17, 18	✓	✓*	✓

✓: sufficient ✓: necessary and sufficient

*: including CPDAGs with background knowledge

Overview of graphical criteria for more general graphs

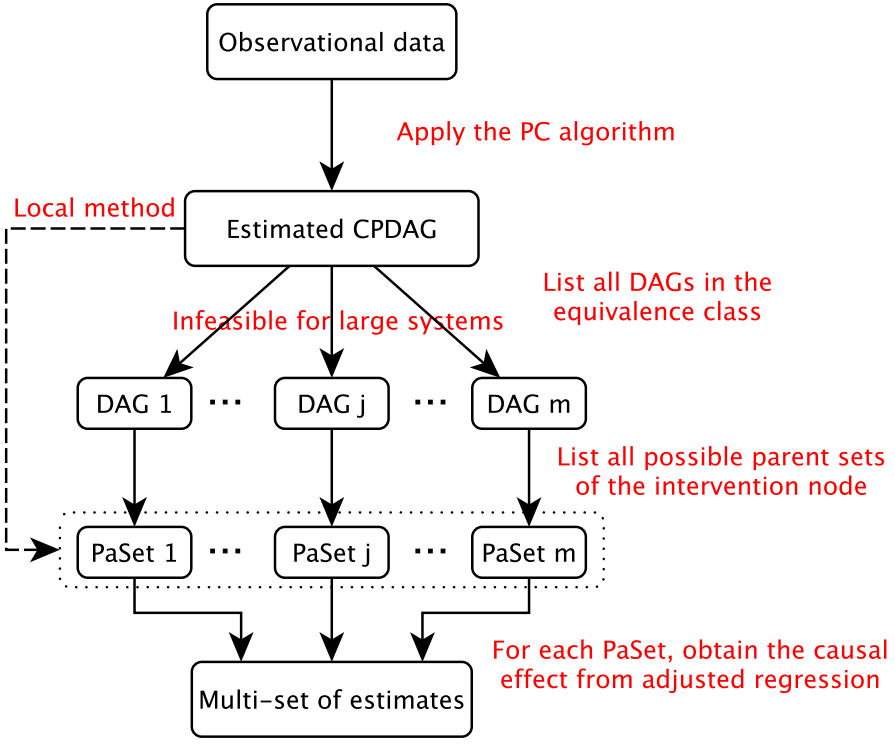
	DAG	CPDAG	PAG
Backdoor criterion Pearl '93	✓		
Adjustment criterion Shpitser et al '12, Perković et al '18	✓		
Generalized backdoor criterion MM & Colombo '15	✓	✓	✓
Generalized adjustment criterion Perković et al '15, '17, 18	✓	✓*	✓

✓: sufficient ✓: necessary and sufficient

*: including CPDAGs with background knowledge

O-set was also generalized to CPDAGs* (Henckel et al '19)

Going away from identifiability: Intervention-calculus when the DAG is Absent (IDA)



Variations of IDA

- Assuming no latent variables:
  - IDA (MM et al '09,'10)
  - joint-IDA: multiple simultaneous interventions (Nandy et al '17)
  - optimal-IDA (Witte et al '20)
- Allowing arbitrarily many latent variables:
  - LV-IDA (Malinsky & Spirtes '17)
- Allowing some latent variables:
  - LRpS-GES IDA (Frot et al '19)

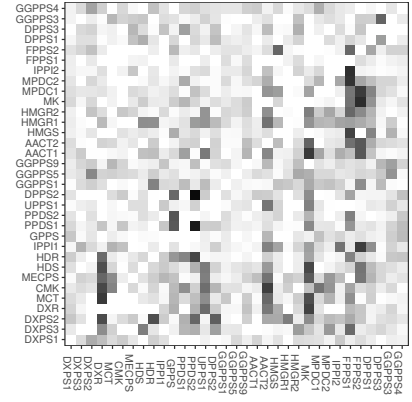
Application

Gene expression data of *Arabidopsis thaliana*:

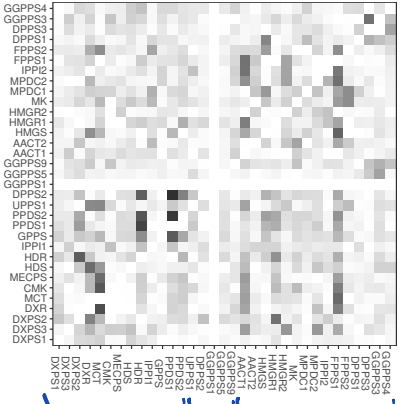
- Data: $n = 188$, $p = 33$ (Wille et al '04)
- Three groups of genes:
  MVA pathway, MEP pathway, mitochondrial genes
  (we did not use this information)
- Goal: estimate lower bounds on the causal effects between all possible gene pairs

Arabidopsis results

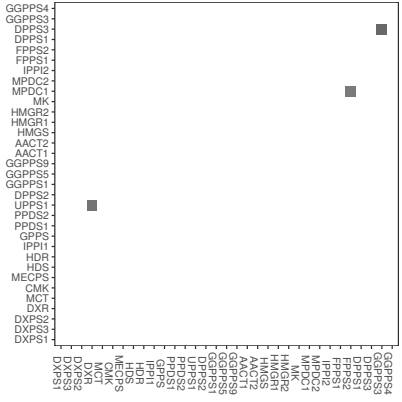
GES-IDA



LRpS-GES-IDA



LV-IDA



Summary

- The variables that are included in a model matter
- If interested in causal effects from observational data:
  - state causal assumptions (e.g., draw DAG)
  - use causal methods (e.g., graphical criteria for covariate adjustment)

Summary

- The variables that are included in a model matter
- If interested in causal effects from observational data:
  - state causal assumptions (e.g., draw DAG)
  - use causal methods (e.g., graphical criteria for covariate adjustment)
- This does not replace randomized controlled trials, but:
  - it uses observational data in a principled way
  - it allows formal discussion
  - it allows sensitivity analysis wrt different causal assumptions
  - if possible, follow-up with validation experiments

Outlook

There are many other interesting connections between causal reasoning and machine learning:

- Robustness & generalizability
- Fairness
- Explainable & interpretable AI
- Reinforcement learning
- Personalized medicine
- ...

and thanks to my collaborators, colleagues, friends and family!

http://stat.math.ethz.ch/∼maathuis
R-packages `pcalg` and `dagitty`

References on causal structure learning

- Spirtes, Glymour and Scheines (2000). *Causation, Prediction, and Search.* MIT Press, Cambridge.
- Chickering (2002). Learning equivalence classes of Bayesian-network structures. *JMLR*.
- Tsamardinos, Brown and Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *JMLR*.
- Colombo, Maathuis, Kalisch and Richardson (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.*
- Claassen, Mooij and Heskes (2013). Learning sparse causal models is not NP-hard. *UAI 2013*.
- Colombo and Maathuis (2014). Order-independent constraint-based causal structure learning. *JMLR*.
- Nandy, Hauser and Maathuis (2018). High-dimensional consistency in score-based and hybrid structure learning. *Ann. Stat.*
- Frot, Nandy and Maathuis (2019). Robust causal structure learning with some hidden variables. *JRSS-B*.
- Eigenmann, Mukherjee and Maathuis (2020). Evaluation of Causal Structure Learning Algorithms via Risk Estimation. *UAI 2020*.

References on adjustment

- Pearl (1993). Comment: Graphical models, causality and intervention. *Stat. Sci.*
- Shpitser, VanderWeele and Robins. On the validity of covariate adjustment for estimating causal effects. *UAI 2010*.
- Van der Zander, Liśkiewicz and Textor. Constructing separators and adjustment sets in ancestral graphs. *UAI 2014*.
- Maathuis and Colombo (2015). A generalized back-door criterion. *Ann. Stat.*
- Perković, Textor, Kalisch and Maathuis (2015). A complete generalized adjustment criterion. *UAI 2015*.
- Perković, Kalisch and Maathuis (2017). Interpreting and using CPDAGs with background knowledge. *UAI 2017*.
- Perković, Textor, Kalisch and Maathuis (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *JMLR*.
- Henckel, Perković and Maathuis (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. arXiv:1907.02435.
- Rotnitzky and Smucler (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *JMLR*.
- Witte, Henckel, Maathuis and Didelez (2020). On efficient adjustment in causal graphs. arXiv:2002:06825.

References on causal structure learning + adjustment

- Maathuis, Kalisch and Bühlmann (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Stat.*
- Maathuis, Colombo, Kalisch and Bühlmann (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*.
- Stekhoven, Moraes, Sveinbjörnsson, Hennig, Maathuis and Bühlmann (2012). Causal stability ranking. *Bioinformatics*.
- Colombo and Maathuis (2014). Order-independent constraint-based causal structure learning. *JMLR*.
- Malinsky and Spirtes (2017). Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *Int. J. Appr. Reason.*
- Perković, Kalisch and Maathuis (2017). Interpreting and using CPDAGs with background knowledge. *UAI 2017*.
- Nandy, Maathuis and Richardson (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Stat.*
- Frot, Nandy and Maathuis (2019). Robust causal structure learning with some hidden variables. *JRSS-B*.
- Witte, Henckel, Maathuis and Didelez (2020). On efficient adjustment in causal graphs. arXiv:2002:06825.