Identifying causal effects from observational data

Emilija Perković
University of Washington

Goal

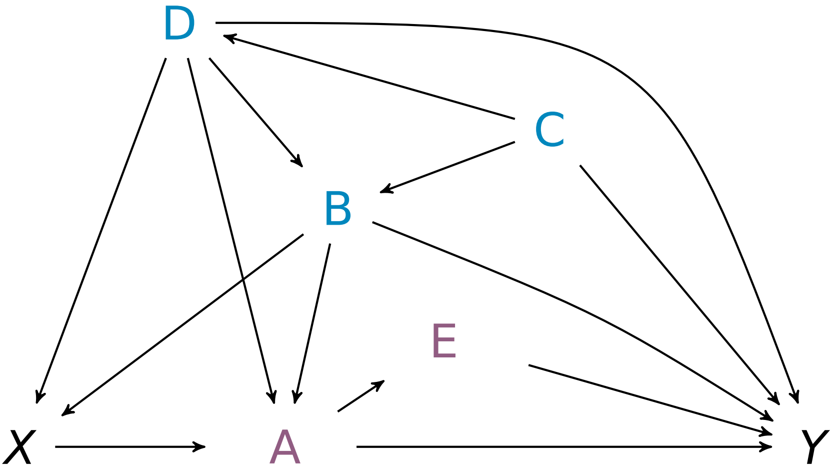
- Estimate the total causal effect of **X** on **Y**
  - the change in **Y** due to $do(\mathbf{x})$-
  from observational data.

- $do(\mathbf{x})$: an intervention that sets variables **X** to **x**.
  $f(\mathbf{y}|do(\mathbf{x})) \neq f(\mathbf{y}|\mathbf{x})$.

Observational data

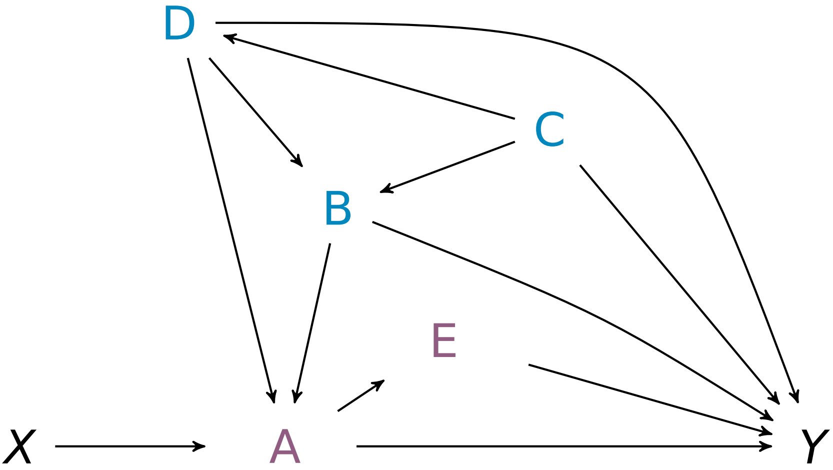
Randomized control studies

Observational Causal DAG



Causal Directed Acyclic Graph (DAG) $\mathcal{D}$.

Interventional Causal DAG



Causal DAG $\mathcal{D}$ **after a "do"-intervention on** X_1**.**

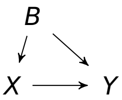
DAGs and distributions

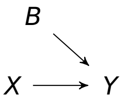
Interventional density

- $do(\mathbf{x})$: an intervention that sets variables $\mathbf{X}$ to $\mathbf{x}$.
- Observational density $f(\mathbf{v})$, Interventional density $f(\mathbf{v}|do(\mathbf{x}))$.

- A DAG $\mathcal{D}$ is causal if for all observational and interventional densities:

$$f(\mathbf{v}) = \prod_{V_j \in \mathbf{V}} f(v_j|pa(v_j,\mathcal{D})) \quad \text{and} \quad f(\mathbf{v}|do(\mathbf{x})) = \prod_{V_j \in \mathbf{V}\setminus\mathbf{X}} f(v_j|pa(v_j,\mathcal{D})).$$





$$f(b,x,y) = f(y|b,x)f(x|b)f(b)$$

$$f(b,y|do(x)) = f(y|b,x)f(b)$$

$$f(b,y|x) = f(y|b,x)f(b|x) \neq f(b,y|do(x))$$

How to define a causal effect?

Total causal effect

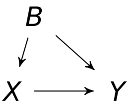
- Total causal effect - $\tau_{\mathbf{yx}}$ - is *some functional* of $f(\mathbf{y}|do(\mathbf{x}))$, $P(\mathbf{Y}|do(\mathbf{x}))$.
- Examples: $E[Y|do(X=x+1)] - E[Y|do(X=x)]$, $\frac{\partial}{\partial x}E(Y|do(x))$, OR, RR...

Identifiability

- A causal effect is identifiable from observational data if

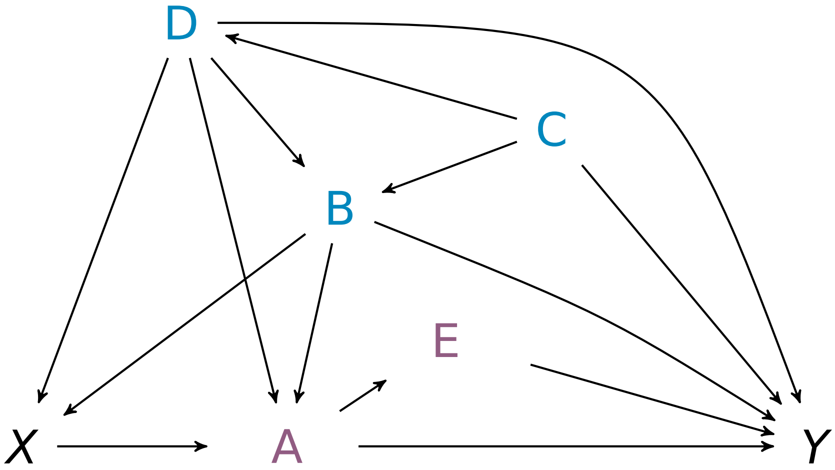
$$f(\mathbf{y}|do(\mathbf{x})) \text{ is computable from } f(\mathbf{v}).$$

- Given the causal DAG, every total causal effect is identifiable.



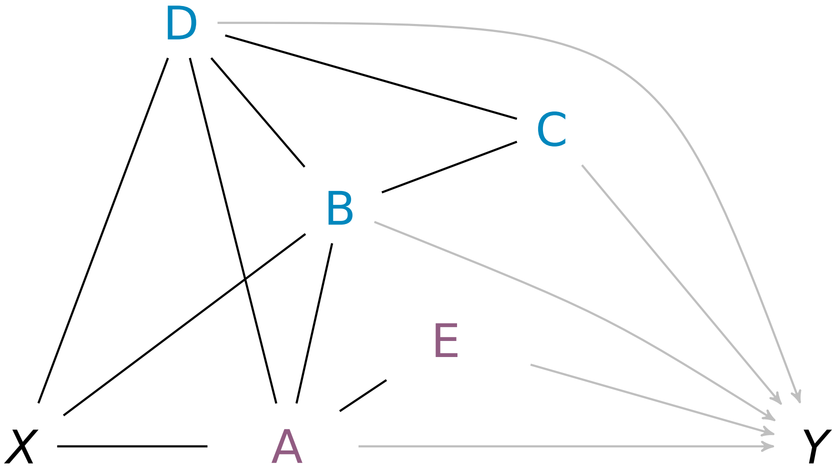
$$f(y|do(x)) = \int f(b,y|do(x))db$$
$$= \int f(y|b,x)f(b)db.$$

Problem solved?



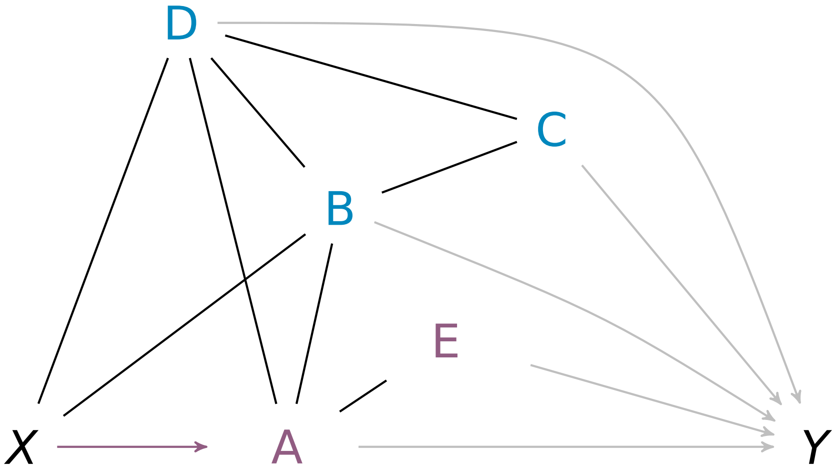
DAG $\mathcal{D}$.

Problem solved?



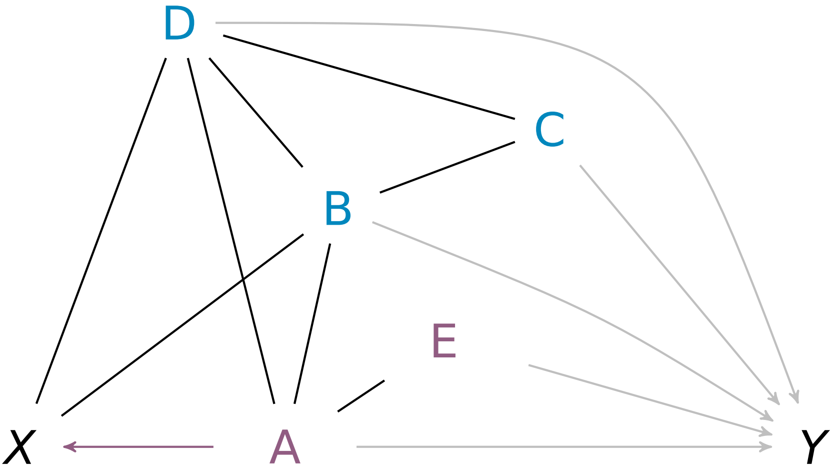
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



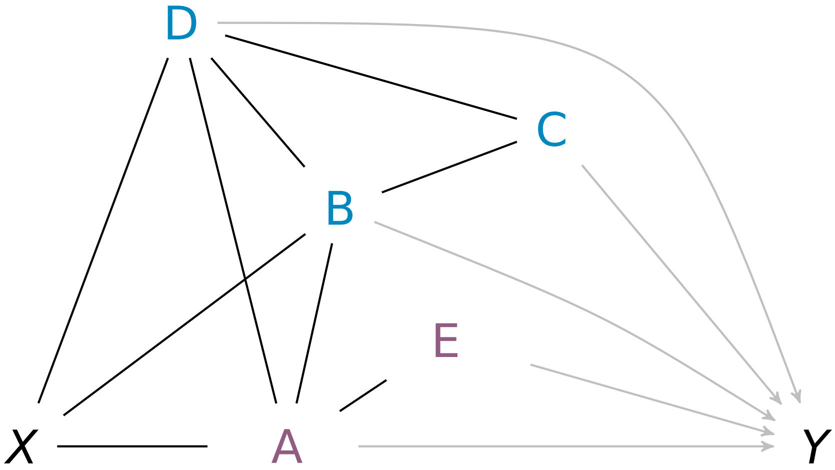
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



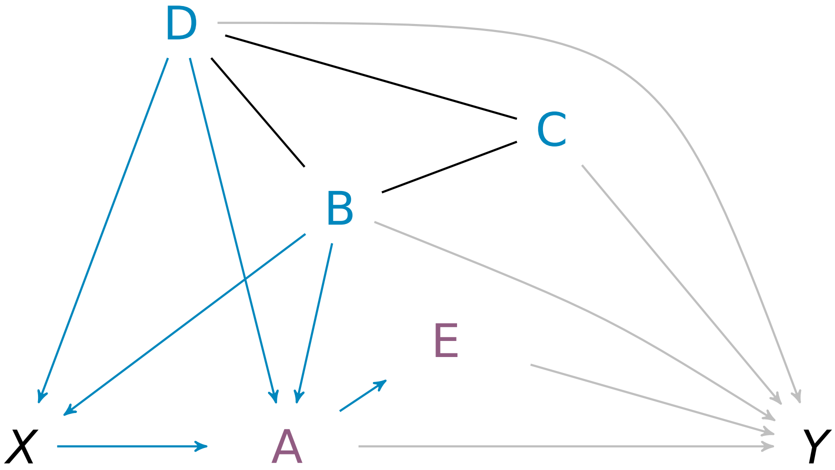
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



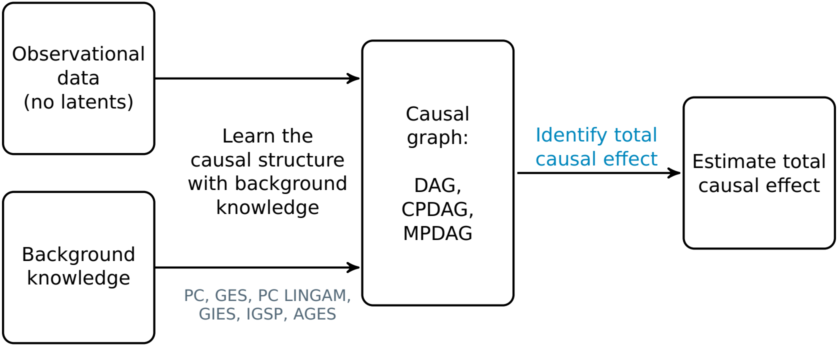
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



Maximally oriented Partially Directed Acyclic Graph (MPDAG) $\mathcal{G}$.

Framework



- PC (Spirtes et al, 1993), GES (Chickering, 2002)
- Adding background knowledge (Meek, 1995; TETRAD, Scheines et al., 1998), PC LINGAM (Hoyer et al., 2008), GIES (Hauser and Bühlmann, 2012), IGSP (Wang et al., 2017), etc.

Overview of graphical criteria for identification

Graphical criterion	DAG	CPDAG	MPDAG
Generalized adjustment (Shpitser et al ’10, Perković et al ’15, ’17, ’18)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
G-formula (Robins ’86)	$\Leftrightarrow$		

$\Rightarrow$ - sufficient for identification,
$\Leftrightarrow$ - necessary and sufficient for identification

Adjustment: $\mathbf{Z}$ is an adjustment set if

$$f(\mathbf{y}|do(\mathbf{x})) = \int f(\mathbf{y}|\mathbf{x},\mathbf{z})f(\mathbf{z})d\mathbf{z}$$

G-formula: Let $\mathbf{V}' = \mathbf{V} \setminus \{\mathbf{X} \cup \mathbf{Y}\}$, then

$$f(\mathbf{y}|do(\mathbf{x})) = \int \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} f(v_i|pa(v_i,\mathcal{D}))d\mathbf{v}'.$$

Does an adjustment set always exist?

If $\mathbf{X} = \{X\}$, $\mathbf{Y} = \{Y\}$:

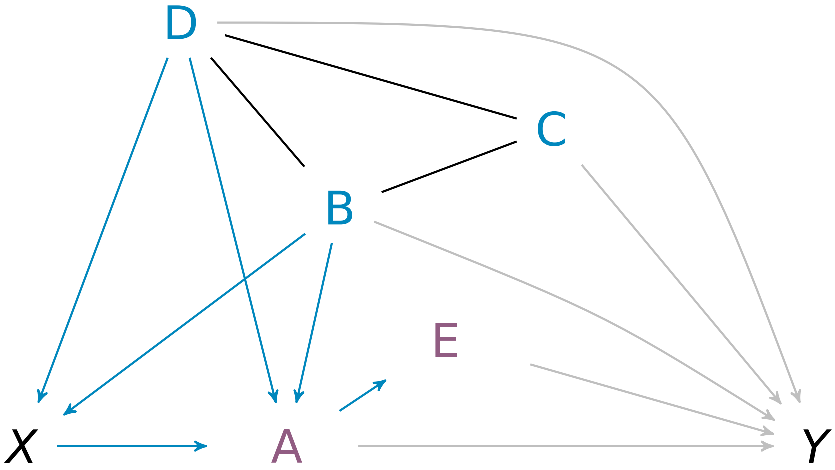
Proposition (Perković, 2020)

If $Y \notin Pa(X, \mathcal{G})$, then an adjustment set relative to (X, Y) exists in the MPDAG $\mathcal{G}$, if and only if the $f(y|do(x))$ is identifiable given $\mathcal{G}$.

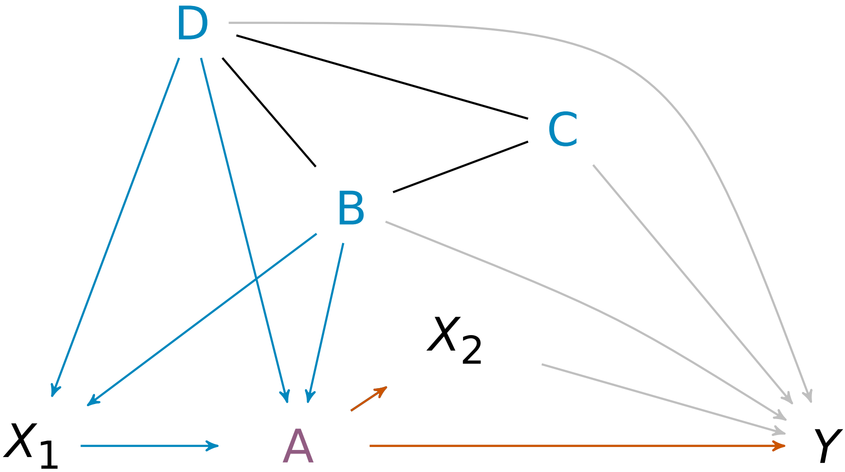
What about for $|\mathbf{X}| > 1$, or $|\mathbf{Y}| > 1$?
Does an adjustment set always exist?

No. Not even in a DAG.

Joint Intervention



Joint intervention



Overview of graphical criteria for identification

Graphical criterion	DAG	CPDAG	MPDAG
Generalized adjustment (Shpitser et al ’10, Perković et al ’15, ’17, ’18)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
G-formula (Robins ’86)	$\Leftrightarrow$		
Causal identification formula (Perković ’20)	$\Leftrightarrow$	$\Leftrightarrow$	$\Leftrightarrow$

$\Rightarrow$ - sufficient for identification,

$\Leftrightarrow$ - necessary and sufficient for identification

Causal identification formula

Theorem (Perković, 2020)

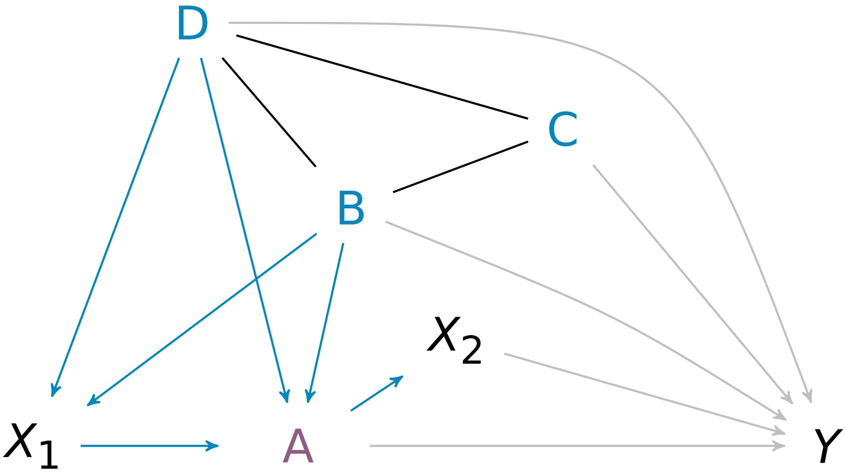
If **all proper possibly causal paths** from $\mathbf{X}$ to $\mathbf{Y}$ start with a directed edge in $\mathcal{G}$, then

$$f(\mathbf{y}|do(\mathbf{x})) = \int \prod_{i=1}^{k} f(\mathbf{s_i}|pa(\mathbf{s_i}, \mathcal{G}))d\mathbf{s},$$

where $\mathbf{S} = an(\mathbf{Y}, \mathcal{G}_{\mathbf{V} \setminus \mathbf{X}}) \setminus \mathbf{Y}$,
and $(\mathbf{S_1}, \dots, \mathbf{S_k})$ is a partition of $\mathbf{S} \cup \mathbf{Y}$ into undirected connected sets in $\mathcal{G}$.

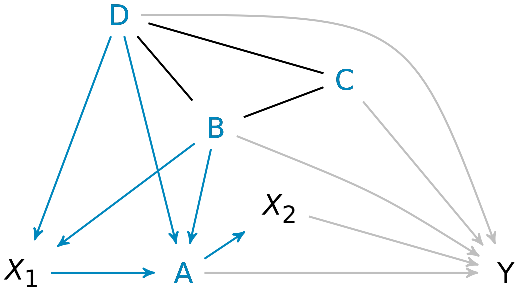
- $\mathbf{S} \cup \mathbf{Y} = an(\mathbf{Y}, \mathcal{G}_{\mathbf{V} \setminus \mathbf{X}})$ - nodes that have a causal path to $\mathbf{Y}$ that is not through $\mathbf{X}$.
- $(\mathbf{S_1}, \dots, \mathbf{S_k})$ - maximal connected components of $\mathbf{S} \cup \mathbf{Y}$ in the induced undirected subgraph of $\mathcal{G}$.

How to use the causal identification formula?



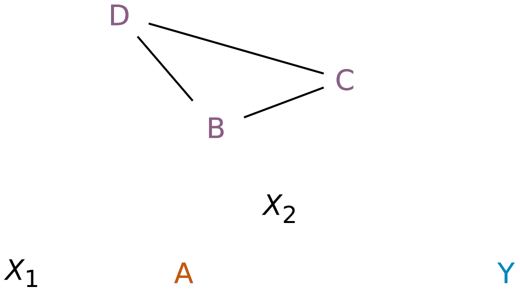
$$f(y|do(x_1,x_2)) = \int f(y|a,b,c,d,x_2)f(a|b,d,x_1)f(b,c,d)\,da\,db\,dc\,dd$$

$f(y|do(x_1, x_2)) = ?$



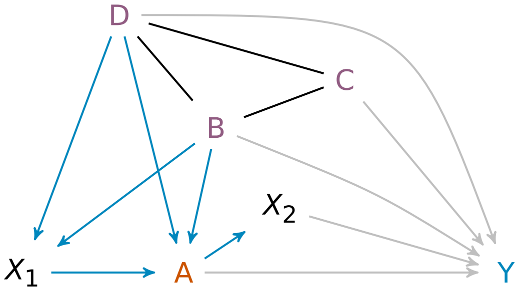
- $\mathbf{S} = an(Y, \mathcal{G}_{\mathbf{V}\setminus\mathbf{X}}) \setminus \{Y\} = \{A, B, C, D\}$, Partition of $\mathbf{S} \cup \{Y\} =$

$f(y|do(x_1, x_2)) = ?$



- $\mathbf{S} = an(Y, \mathcal{G}_{\mathbf{V} \setminus \mathbf{X}}) \setminus \{Y\} = \{A, B, C, D\}$, Partition of $\mathbf{S} \cup \{Y\} = (\{B, C, D\}, \{A\}, \{Y\})$.

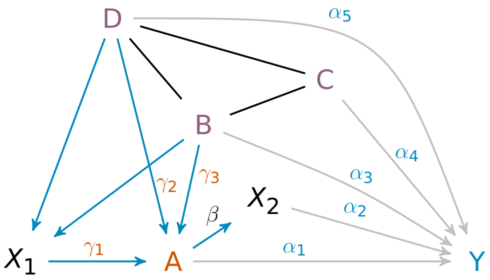
$f(y|do(x_1, x_2)) = ?$



- $\mathbf{S} = an(Y, \mathcal{G}_{\mathbf{V}\setminus\mathbf{X}}) \setminus \{Y\} = \{A, B, C, D\}$, Partition of $\mathbf{S} \cup \{Y\} = (\{B, C, D\}, \{A\}, \{Y\})$.

$$
\begin{aligned}
f(y|do(x_1, x_2)) &= \int f(y, \mathbf{s}|do(x_1, x_2))d\mathbf{s} = \int f(y, a, b, c, d|do(x_1, x_2))d\mathbf{s} \\
&= \int f(y|a, b, c, d, do(x_1, x_2))f(a|b, c, d, do(x_1, x_2))f(b, c, d|do(x_1, x_2))d\mathbf{s} \\
&= \int f(y|a, b, c, d, do(x_1, x_2))f(a|b, d, do(x_1, x_2))f(b, c, d|do(x_1, x_2))d\mathbf{s} \\
&= \int f(y|a, b, c, d, x_2)f(a|b, d, x_1)f(b, c, d)d\mathbf{s}.
\end{aligned}
$$

Estimation in the linear case



- $\tau_{y\mathbf{x}} = (\tau_{yx_1.x_2}, \tau_{yx_2.x_1})^T = (\alpha_1\gamma_1, \alpha_2)^T = \left(\frac{\partial E[Y|do(x_1,x_2)]}{\partial x_1}, \frac{\partial E[Y|do(x_1,x_2)]}{\partial x_2}\right)^T$

$$E[Y|do(x_1,x_2))] = \int yf(y|a,b,c,d,x_2)f(a|b,d,x_1)f(b,c,d)d\mathbf{s}$$

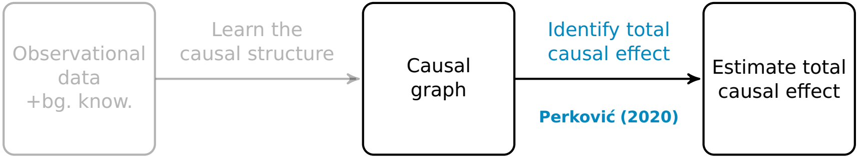
$$= \int E[Y|a,b,c,d,x_2]f(a|b,d,x_1)f(b,c,d)d\mathbf{s}$$

$$= \int (\alpha_1 a + \alpha_2 x_2 + \alpha_3 b + \alpha_4 c + \alpha_5 d)f(a|b,d,x_1)f(b,c,d)d\mathbf{s}$$

$$= \alpha_1 \int E[A|b,d,x_1]f(b,d)db\,dd + \alpha_2 x_2 + \int (\alpha_3 b + \alpha_4 c + \alpha_5 d)f(b,c,d)db\,dc\,dd$$

$$= \alpha_1\gamma_1 x_1 + \alpha_2 x_2 + (\alpha_1\gamma_3 + \alpha_3)E[B] + \alpha_4 E[C] + (\alpha_1\gamma_2 + \alpha_5)E[D].$$

Summary



- Graphical necessary condition for identification of $f(\mathbf{y}|do(\mathbf{x}))$.
- Necessary and sufficient graphical criterion for identification of causal effects.
- Proposition on the "necessity" of adjustment.

Thanks!