Introduction to Potential Outcomes

Thomas Richardson
University of Washington

Simons Causal Bootcamp
20 January 2022

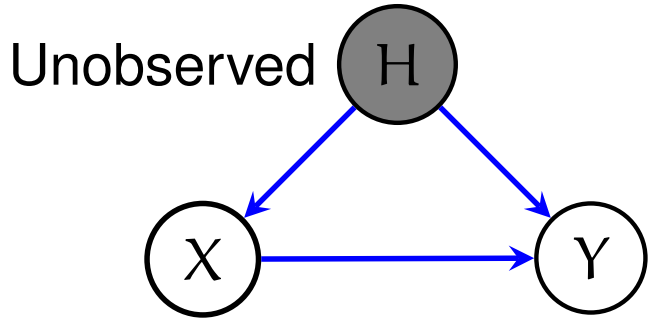
Joint work with James M. Robins (Harvard) and Ilya Shpitser (JHU)

Outline

- Part One: SWIGs
  - Relating graphs and counterfactuals via node-splitting
  - Simple examples
  - General procedure
  - Factorization and Modularity Properties
  - Adjustment for Confounding
  - Potential Outcomes (po) Calculus
  - Front-door criterion via po-calculus
  - Sequentially Randomized Experiments / Time Dependent Confounding
- Part Two: Mediation
  - Separable Direct Effects
  - Contrast with other approaches: Controlled, Principal Stratum and Pure (aka Natural) Direct Effect

Graphical Approach to Causality

No Confounding



Confounding

- Graph intended to represent direct causal relations.
- Convention that confounding variables (e.g. H) are always included on the graph.
- Approach originates in the path diagrams introduced by Sewall Wright in the 1920s.
- If $X \to Y$ then X is said to be a *parent* of Y; Y is *child* of X.

Graphical Approach to Causality

No Confounding

- Associated factorization:

$$P(x, y) = P(x)P(y \mid x)$$

- In the absence of confounding the *causal* model asserts:

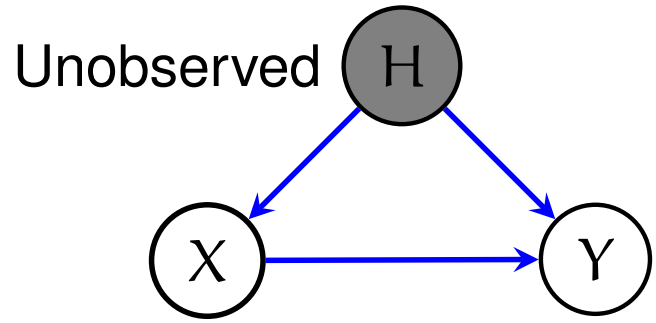
$$P(Y(x) = y) = P(Y = y \mid \mathrm{do}(X = x)) = P(Y = y \mid X = x).$$

  Thus $\mathrm{ACE}(X \to Y)$ is identified under this model.

- Q: *How does this relate to the non-graphical approach?*

Linking the two approaches

$X \perp\!\!\!\perp Y(x_0)$ & $X \perp\!\!\!\perp Y(x_1)$

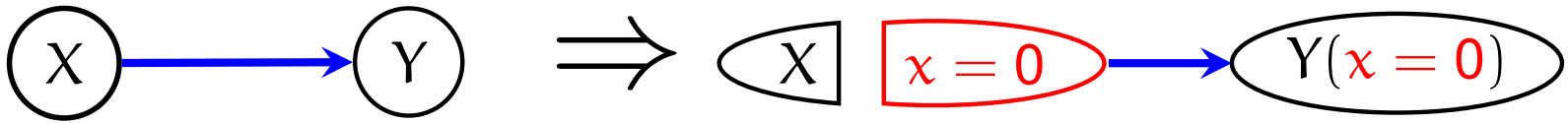


$X \not\perp\!\!\!\perp Y(x_0)$ *or* $X \not\perp\!\!\!\perp Y(x_1)$

- Elephant in the room:
  The variables $Y(x_0)$ *and* $Y(x_1)$ *do not appear on these graphs!!*

Node splitting: Setting X to 0

$$P(X=\tilde{x}, Y=\tilde{y}) = P(X=\tilde{x})P(Y=\tilde{y} \mid X=\tilde{x})$$



Can now 'read' the independence: $X \perp\!\!\!\perp Y(x=0)$.
Also associate a new factorization:

$$P\left(X=\tilde{x}, Y(x=0)=\tilde{y}\right) = P(X=\tilde{x})P\left(Y(x=0)=\tilde{y}\right)$$

where:

$$P\left(Y(x=0)=\tilde{y}\right) = P(Y=\tilde{y} \mid X=0).$$

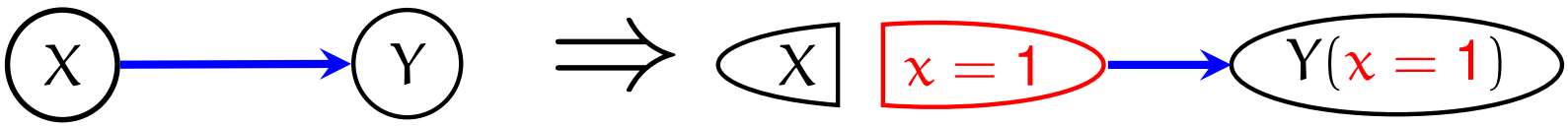
This last equation links a term in the original factorization to the new factorization. We term this the 'modularity assumption'.

From counterfactual perspective modularity follows from factorization + consistency:

$$P\left(Y(x=0)=\tilde{y}\right) = P\left(Y(x=0)=\tilde{y} \mid X=0\right) = P(Y=\tilde{y} \mid X=0)$$

Node splitting: Setting X to 1

$$P(X=\tilde{x}, Y=\tilde{y}) = P(X=\tilde{x})P(Y=\tilde{y} \mid X=\tilde{x})$$



Can now 'read' the independence: $X \perp\!\!\!\perp Y(x=1)$.
Also associate a new factorization:

$$P\left(X=\tilde{x}, Y(x=1)=\tilde{y}\right) = P(X=\tilde{x})P\left(Y(x=1)=\tilde{y}\right)$$

where:

$$P\left(Y(x=1)=y\right) = P(Y=y \mid X=1).$$

Marginals represented by SWIGs are identified

The SWIG $\mathcal{G}(x_0)$ represents $P(X, Y(x_0))$.
The SWIG $\mathcal{G}(x_1)$ represents $P(X, Y(x_1))$.
Under no confounding these marginals are identified from $P(X, Y)$.
In contrast the distribution $P(X, Y(x_0), Y(x_1))$ is not identified.
$Y(x=0)$ and $Y(x=1)$ are never on the same SWIG.
Although we have:

$$X \perp\!\!\!\perp Y(x=0) \quad \text{and} \quad X \perp\!\!\!\perp Y(x=1)$$

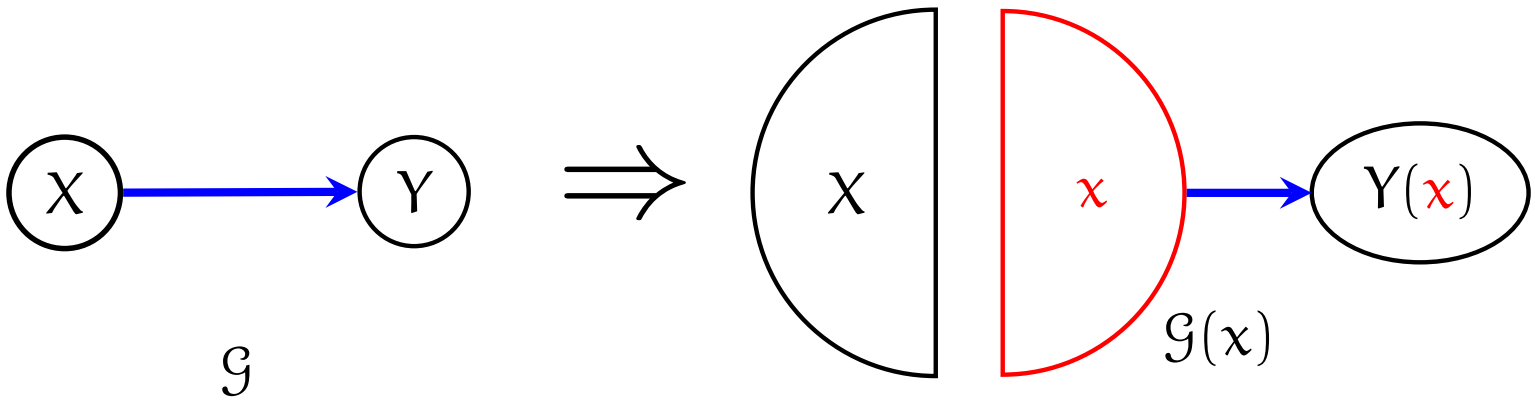
we do not assume

$$X \perp\!\!\!\perp Y(x=0), Y(x=1)$$

Had we tried to construct a single graph containing both $Y(x=0)$ and $Y(x=1)$ this would have been impossible.

$\Rightarrow$ *Single-World Intervention Graphs* (SWIGs).

Representing both graphs via a 'template'



Represent both graphs via a *template*:

Formally the template is a 'graph valued function' (not a graph!):

- Takes as input a specific value x^*
- Returns as output a SWIG $\mathcal{G}(x^*)$.

Each *instantiation* of the template represents a different margin:
SWIG $\mathcal{G}(x_0)$ represents $P(X, Y(x_0))$;
SWIG $\mathcal{G}(x_1)$ represents $P(X, Y(x_1))$.

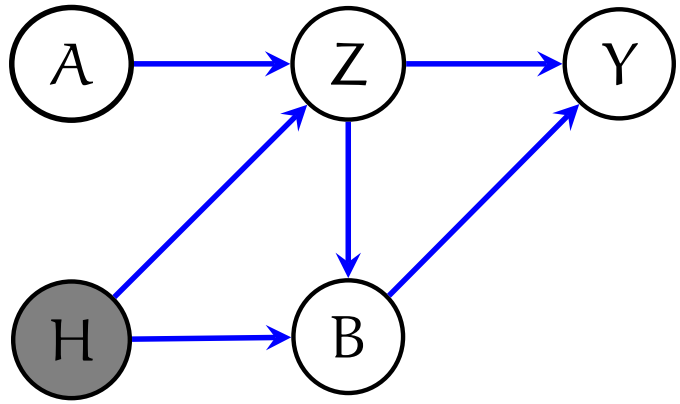
Intuition behind node splitting:

(Robins, VanderWeele, Richardson 2007)

Q: *How could we identify whether someone would choose to take treatment, i.e. have* $X = 1$*, and at the same time find out what happens to such a person if they don't take treatment* $Y(x = 0)$*?*

A: Consider an experiment in which, whenever a patient is observed to swallow the drug have $X = 1$, we instantly intervene by administering a safe 'emetic' that causes the pill to be regurgitated before any drug can enter the bloodstream. Since we assume the emetic has no side effects, the patient's recorded outcome is then $Y(x = 0)$.

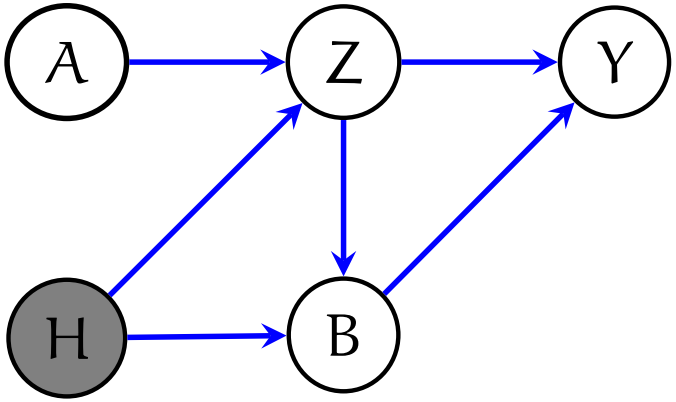
Harder Inferential problem

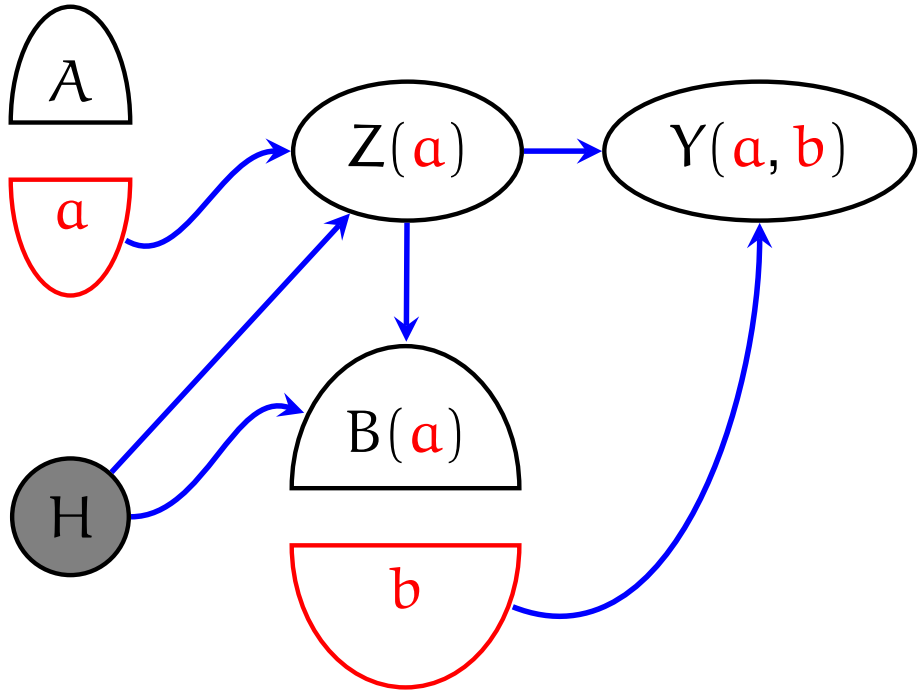


Query: does this causal graph imply:

$$Y(a,b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

Simple solution





Query does this graph imply:

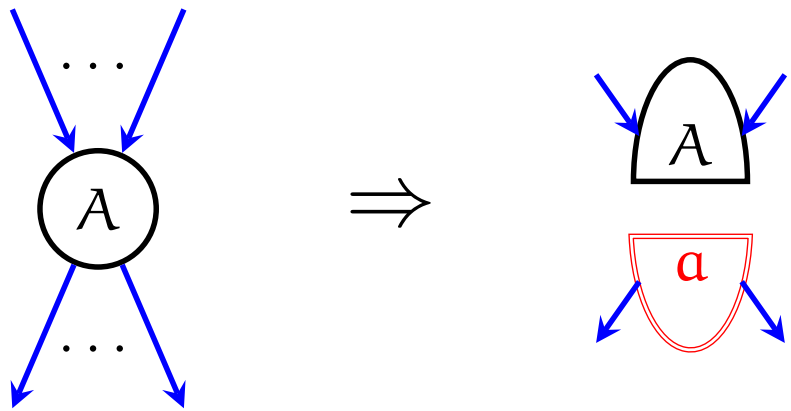
$$Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

Answer: Yes – applying d-separation to the SWIG on the right we see that there is no d-connecting path from $Y(a, b)$ given $Z(a)$.

Single World Intervention Template Construction (1)

Given a graph G, a subset of vertices $\mathbf{A} = \{A_1, \ldots, A_k\}$ to be intervened on, we form $\mathrm{G}(\mathbf{a})$ in two steps:

(1) (Node splitting): For every $A \in \mathbf{A}$ split the node into a *random* node A and a *fixed* node a:

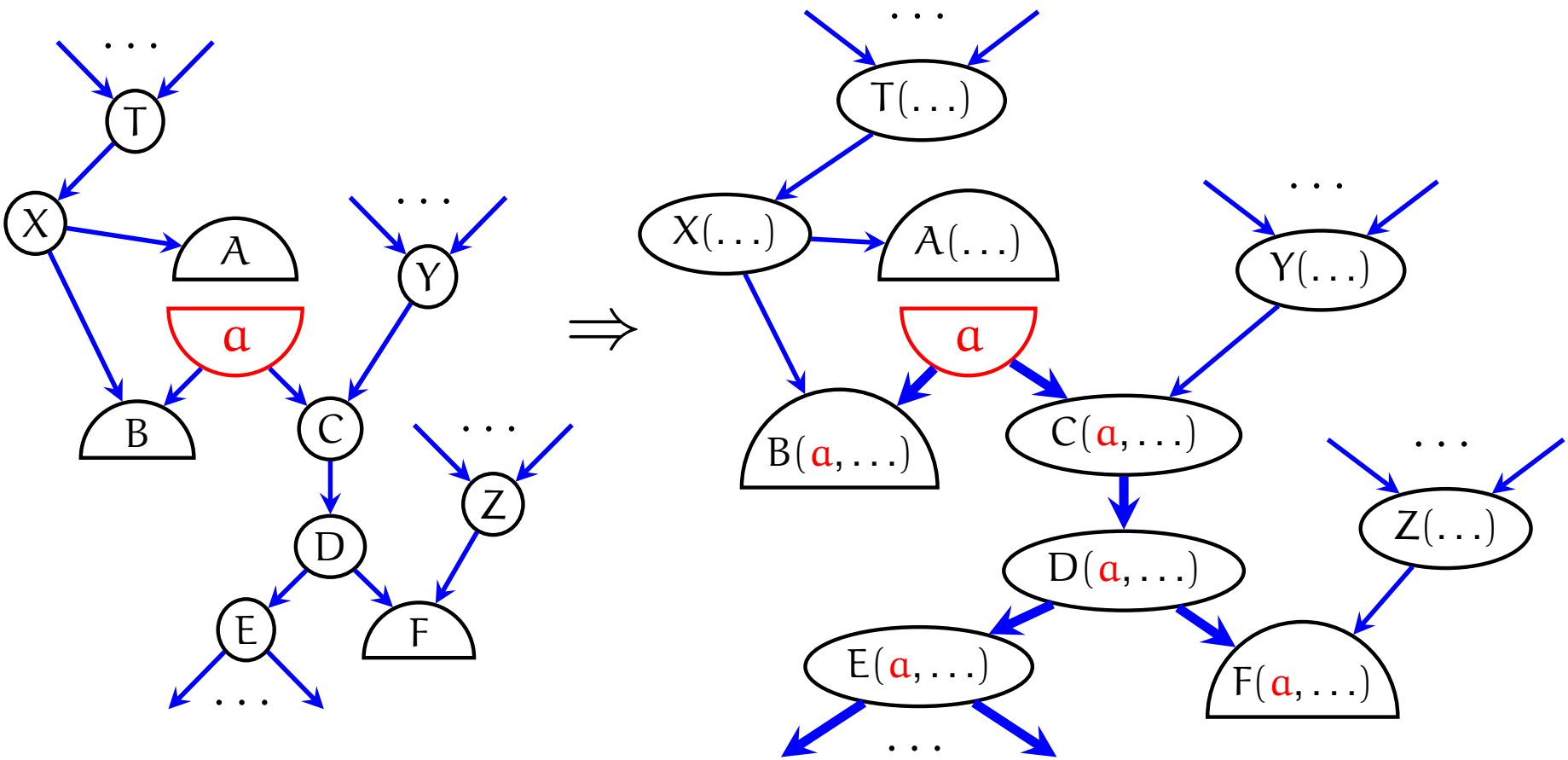


Splitting: Schematic Illustrating the Splitting of Node A

- The random half inherits all edges directed into A in $\mathcal{G}$;
- The fixed half inherits all edges directed out of A in $\mathcal{G}$.

Single World Intervention Template Construction (2)

(2) Relabel descendants of fixed nodes:



Summary Adding Counterfactual Distributions to DAGs

Original graph $\mathcal{G}$: observed distribution $\mathrm{P}(\mathbf{V})$
SWIG $\mathcal{G}(\tilde{\mathbf{a}})$: counterfactual distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$

Note that under minimal labeling variables in $\mathbf{V}(\tilde{\mathbf{a}})$ may be not labelled with the full set $\tilde{\mathbf{a}}$.

Factorization of counterfactual variables: Distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ over the variables in $\mathcal{G}(\tilde{\mathbf{a}})$ factorizes with respect to the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ (ignoring fixed nodes):

Modularity: $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ and $\mathrm{P}(\mathbf{V})$ are linked as follows:
The conditional density associated with $Y(\tilde{\mathbf{a}})$ in $\mathcal{G}(\tilde{\mathbf{a}})$ is just the conditional density associated with Y in $\mathcal{G}$ after substituting $\tilde{a}_i$ for any $A_i \in \mathbf{A}$ that is a parent of Y.

Consequence: if $\mathrm{P}(\mathbf{V})$ is observed then $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ is identified.

Applying d-separation to the graph $\mathrm{G}(\mathbf{a})$ (Part 1)

We extend the definition of d-connection to SWIGs as follows:

- A red (fixed) node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a red (fixed) node can d-connect that node to a random node if it satisfies the usual conditions on colliders and non-colliders;

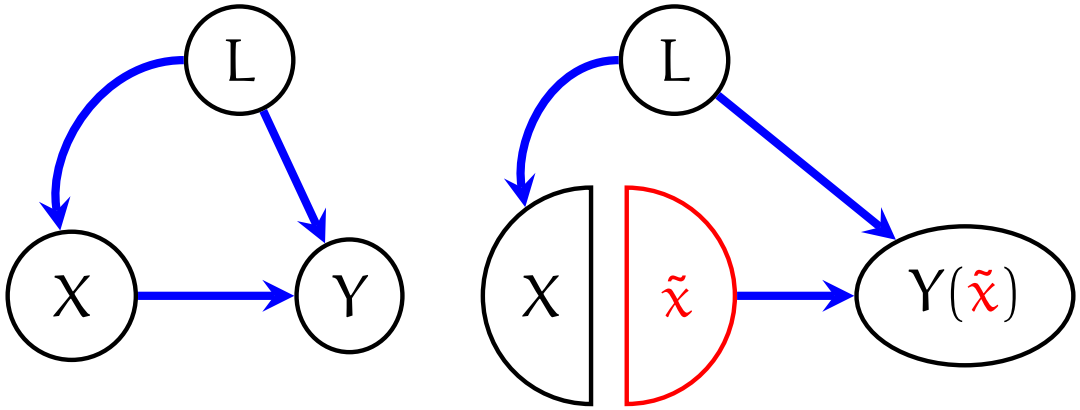
In $\mathcal{G}(\tilde{\mathbf{a}})$ if subsets $\mathbf{B}(\tilde{\mathbf{a}})$ and $\mathbf{C}(\tilde{\mathbf{a}})$ of random nodes are d-separated by $\mathbf{D}(\tilde{\mathbf{a}})$, then $\mathbf{B}(\tilde{\mathbf{a}})$ and $\mathbf{C}(\tilde{\mathbf{a}})$ are conditionally independent given $\mathbf{D}(\tilde{\mathbf{a}})$ in the associated distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$.

$$\mathbf{B}(\tilde{\mathbf{a}}) \text{ is d-separated from } \mathbf{C}(\tilde{\mathbf{a}}) \text{ given } \mathbf{D}(\tilde{\mathbf{a}}) \text{ in } \mathcal{G}(\tilde{\mathbf{a}}) \qquad (1)$$

$$\Rightarrow \quad \mathbf{B}(\tilde{\mathbf{a}}) \perp\!\!\!\perp \mathbf{C}(\tilde{\mathbf{a}}) \mid \mathbf{D}(\tilde{\mathbf{a}}) \quad [\mathbf{P}(\mathbf{V}(\tilde{\mathbf{a}}))].$$

Adjustment for Confounding

Adjusting for confounding

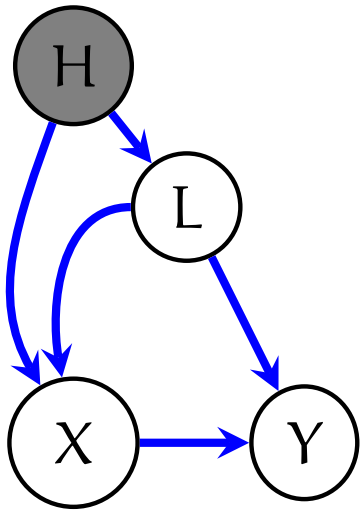


Here we can read directly from the template that

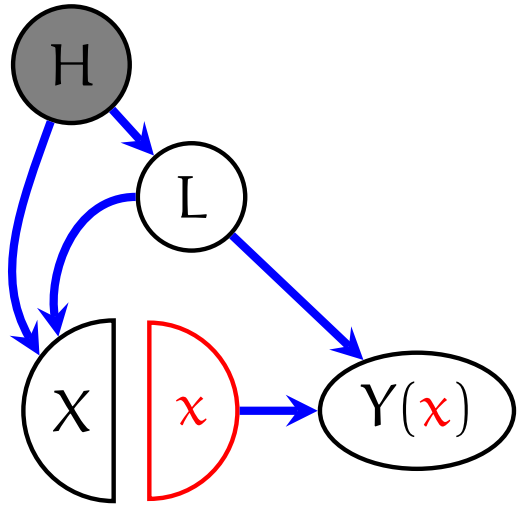
$$X \perp\!\!\!\perp Y(\tilde{x}) \mid L.$$

$$\begin{aligned} P[Y(\tilde{x}) = y] &= \sum_l P[Y(\tilde{x}) = y \mid L = l] P(L = l) \\ &= \sum_l P[Y(\tilde{x}) = y \mid L = l, X = \tilde{x}] P(L = l) \text{ indep} \\ &= \sum_l P[Y = y \mid L = l, X = \tilde{x}] P(L = l) \text{ consistency} \end{aligned}$$

More Examples (I)



(a-i)

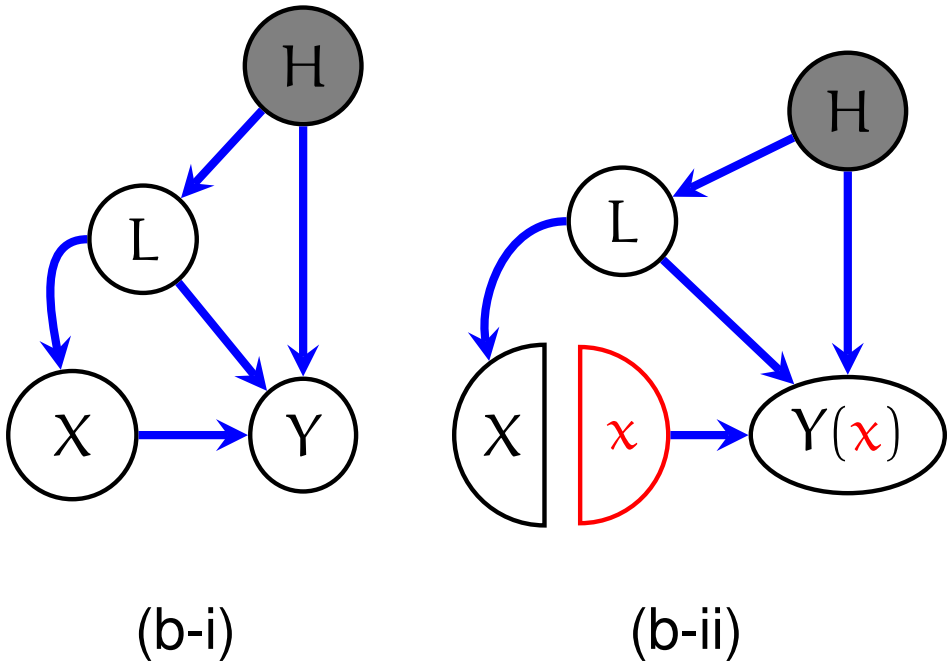


(a-ii)

Here we can read directly from the template that

$$X \perp\!\!\!\perp Y(x) \mid L.$$

More Examples (II)

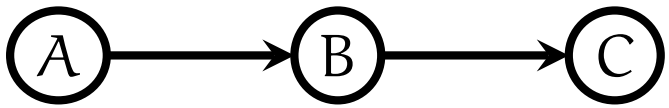


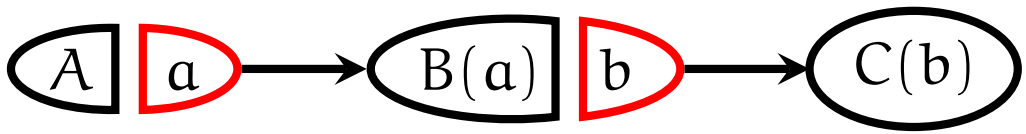
(b-i) (b-ii)

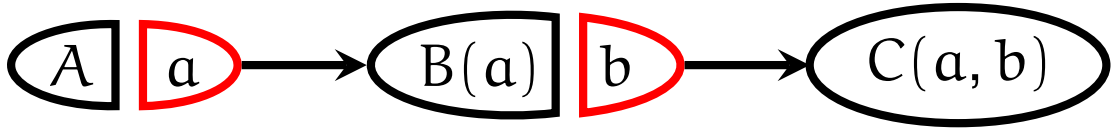
Here we can read directly from the template that

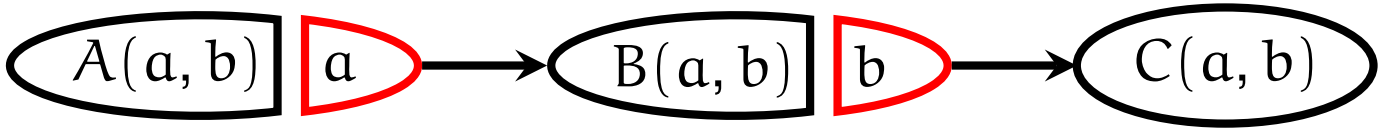
$$X \perp\!\!\!\perp Y(x) \mid L.$$

Labeling Schemes

DAG $\mathcal{G}$ 

Ancestral: 

Temporal: 

Uniform: 

Elsewhere we have used the term ‘minimal’ for the Ancestral Scheme.

Simplifying the *do*-Calculus

Applying d-separation to the graph $\mathrm{G}(\mathbf{a})$ (Part 2)

(Malinsky, Shpitser, R, 2019; Robins 2018)

We extend the definition of d-connection to SWIGs as follows:

- A red (fixed) node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a red (fixed) node can d-connect that node to a random node if it satisfies the usual conditions on colliders and non-colliders;

In $\mathcal{G}(\tilde{\mathbf{a}}, \mathrm{d})$, if fixed node $\mathbf{d}$ is d-separated from $\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d})$ given $\mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d})$ then

$$\mathrm{P}(\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d}) \mid \mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d})) = \mathbf{P}(\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d}') \mid \mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d}')). \tag{2}$$

In other words, the conditional distribution of B given C after intervening on $\mathbf{A}$ and D does not depend on the value assigned to D.

do-calculus

Pearl (1995) formulated a set of rules that give graphical conditions allowing three transformations:

1: Removing observations

$$\begin{aligned} p(y \mid z, w, \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x) \mid Z(x), W(x)) &= p(Y(x) \mid W(x)) \end{aligned}$$

2: Interchanging observation and intervention

$$\begin{aligned} p(y \mid z, w, \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(z), \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x) \mid Z(x), W(x)) &= p(Y(x, z) \mid W(x, z)) \end{aligned}$$

3: Removing interventions:

$$\begin{aligned} p(y \mid w, \mathsf{do}(z), \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x, z) \mid W(x, z)) &= p(Y(x) \mid W(x)) \end{aligned}$$

Do-calculus (details)

Pearl's do-calculus as originally formulated:

$$1: p(y \mid z, w, do(x)) = p(y \mid w, do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X}}}$$

$$2: p(y \mid z, w, do(x)) = p(y \mid w, do(z), do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X},\underline{Z}}}$$

$$3: p(y \mid w, do(z), do(x)) = p(y \mid w, do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$$

where $\mathcal{G}_{\overline{X}}$ denotes the graph obtained from $\mathcal{G}$ by removing all edges with arrowheads into X, $\mathcal{G}_{\underline{Z}}$ denotes the graph obtained from $\mathcal{G}$ by removing all directed edges out of Z, and $Z(W)$ is all elements in Z that are not ancestors of W in $\mathcal{G}_{\overline{X}}$.

Potential Outcomes (PO) Calculus (Malinsky, Shpitser, R, 2019; Shpitser, R, Robins, 2020)

- Can use SWIGs to formulate (simpler, wlog) counterfactual versions of Pearl's rules.

1: If $Y(x)$ is d-separated from $Z(x)$ given $W(x)$ in $\mathcal{G}(x)$ then

$$p(Y(x) \mid Z(x), W(x)) = p(Y(x) \mid W(x))$$

2: If $Y(x, z)$ is d-separated from $Z(x, z)$ given $W(x, z)$ in $\mathcal{G}(x, z)$ then

$$p(Y(x, z) \mid W(x, z)) = p(Y(x) \mid W(x), Z(x) = z)$$

3: If z has no directed path to $Y(x, z)$ in $\mathcal{G}(x, z)$ then

$$p(Y(x, z)) = p(Y(x))$$

- Note: here we use uniform labelings: e.g. $Z(x, z)$ is the random node for Z in $\mathcal{G}(x, z)$. to make explicit which node is in which graph.

Potential Outcomes Calculus: *TL;DR versions*

Suppressing the intervention on X to reduce clutter:

1: If Y is d-separated from Z given W in $\mathcal{G}$ then

$$p(Y \mid Z, W) = p(Y \mid W) \quad \text{(Markov property)}.$$

2: If $Y(z)$ is d-separated from $Z(z)$ given $W(z)$ in $\mathcal{G}(z)$ then

$$p(Y(z) \mid W(z)) = p(Y \mid W, Z = z) \quad \text{(generalized ignorability)}.$$

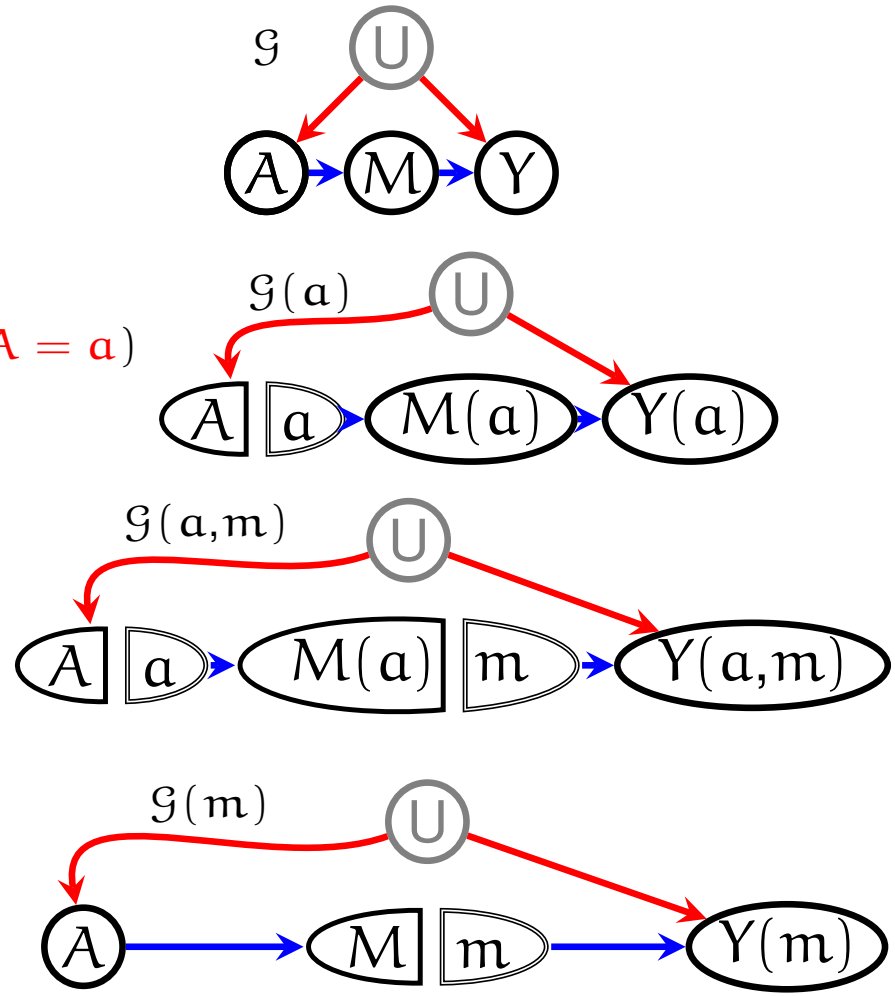
3: If z has no directed path to $Y(z)$ in $\mathcal{G}(z)$ then

$$p(Y(z)) = p(Y) \quad \text{(causal irrelevance)}.$$

po-calculus $=$ d-separation $+$ ignorability
$+$ interventions only affect causal descendants
(!)

Example Derivation (Front-Door)

$$
\begin{aligned}
& p(Y(a)) \\
&= ^{p} \sum_{m} p(Y(a)|M(a)=m)p(M(a)=m) \\
&= ^{2,\mathcal{G}(a)} \sum_{m} p(Y(a)|M(a)=m)p(M=m|A=a) \\
&= ^{2,\mathcal{G}(a,m)} \sum_{m} p(Y(a,m))p(m|a) \\
&= ^{3,\mathcal{G}(a,m)} \sum_{m} p(Y(m))p(m|a) \\
&= ^{p} \sum_{m} p(m|a) \sum_{a'} p(Y(m)|a')p(a') \\
&= ^{2,\mathcal{G}(m)} \sum_{m} p(m|a) \sum_{a'} p(Y|m,a')p(a')
\end{aligned}
$$



How is simplification of Rule 3 possible?

Answer: instances of Rule 3 are already implied by Rule 1 and Rule 2:

Recall condition for Pearl's Rule 3: $(Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$
where $\mathcal{G}_{\overline{X}}$ denotes the graph obtained from $\mathcal{G}$ by removing all edges with arrowheads into X, and $Z(W)$ is all elements in Z that are not ancestors of W in $\mathcal{G}_{\overline{X}}$.

Let $Z_1 = Z \cap \text{an}_{\overline{\mathcal{G}}}(W)$ and $Z_2 = Z \setminus \text{an}_{\overline{\mathcal{G}}}(W) \equiv Z(W)$.

If Pearl's rule 3 applies then Z_1 is d-separated from Y given W in $\mathcal{G}_{\overline{X}\overline{Z_2}\underline{Z_1}}$ as this is a subgraph of $\mathcal{G}_{\overline{X}\overline{Z_2}} \equiv \mathcal{G}_{\overline{X}\overline{Z(W)}}$. So

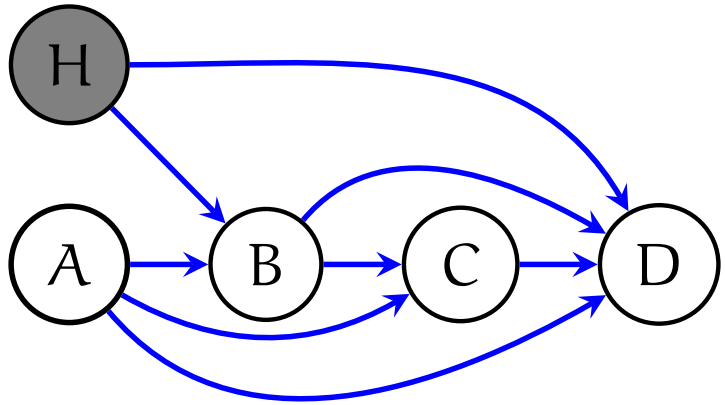
$$\begin{aligned} p(Y \mid do(Z, X), W) &= p(Y \mid do(Z_1, Z_2, X), W) \\ &= p(Y \mid do(Z_2, X), Z_1, W) \quad (\text{Rule2}) \\ &= p(Y \mid do(Z_2, X), W) \quad (\text{Rule1}) \end{aligned}$$

The last step follows since if there were path d-connecting Z_1 and Y given W in $\mathcal{G}_{\overline{X}\overline{Z_2}} \equiv \mathcal{G}_{\overline{X}\overline{Z(W)}}$ then the conditions for Rule 3 does not hold.

$\Rightarrow$ so any vertex in Z that is an ancestor of W may be removed using Rule 1 + 2. Remaining d-conn paths given W take form $Z_2^* \rightarrow \cdots \rightarrow Y$.

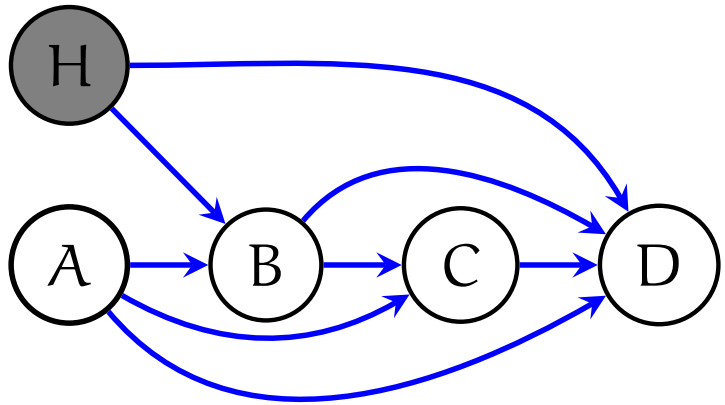
Multiple Treatments

Sequentially randomized experiment (I)



- A and C are treatments;
- H is unobserved (underlying health status);
- B is an initial observed response;
- D is the final response;
- Treatment C is assigned randomly conditional on the observed history, A and B;
- Want to know $\mathrm{P}(\mathrm{D}(\tilde{a}, \tilde{c}))$.

Sequentially randomized experiment (I)



If the following holds:

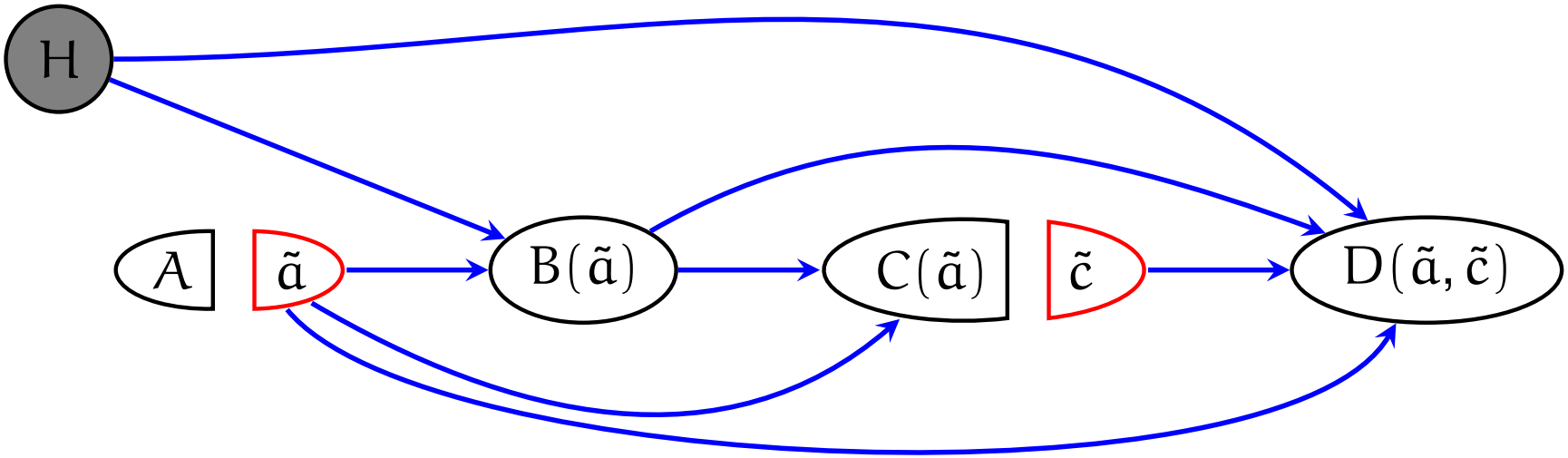
$$\begin{aligned} A &\perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \\ C(\tilde{a}) &\perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \mid B(\tilde{a}), A \end{aligned}$$

General result of Robins (1986) then implies:

$$P(D(\tilde{a}, \tilde{c}) = d) = \sum_b P(B = b \mid A = \tilde{a}) P(D = d \mid A = \tilde{a}, B = b, C = \tilde{c}).$$

Does it??

Sequentially randomized experiment (II)



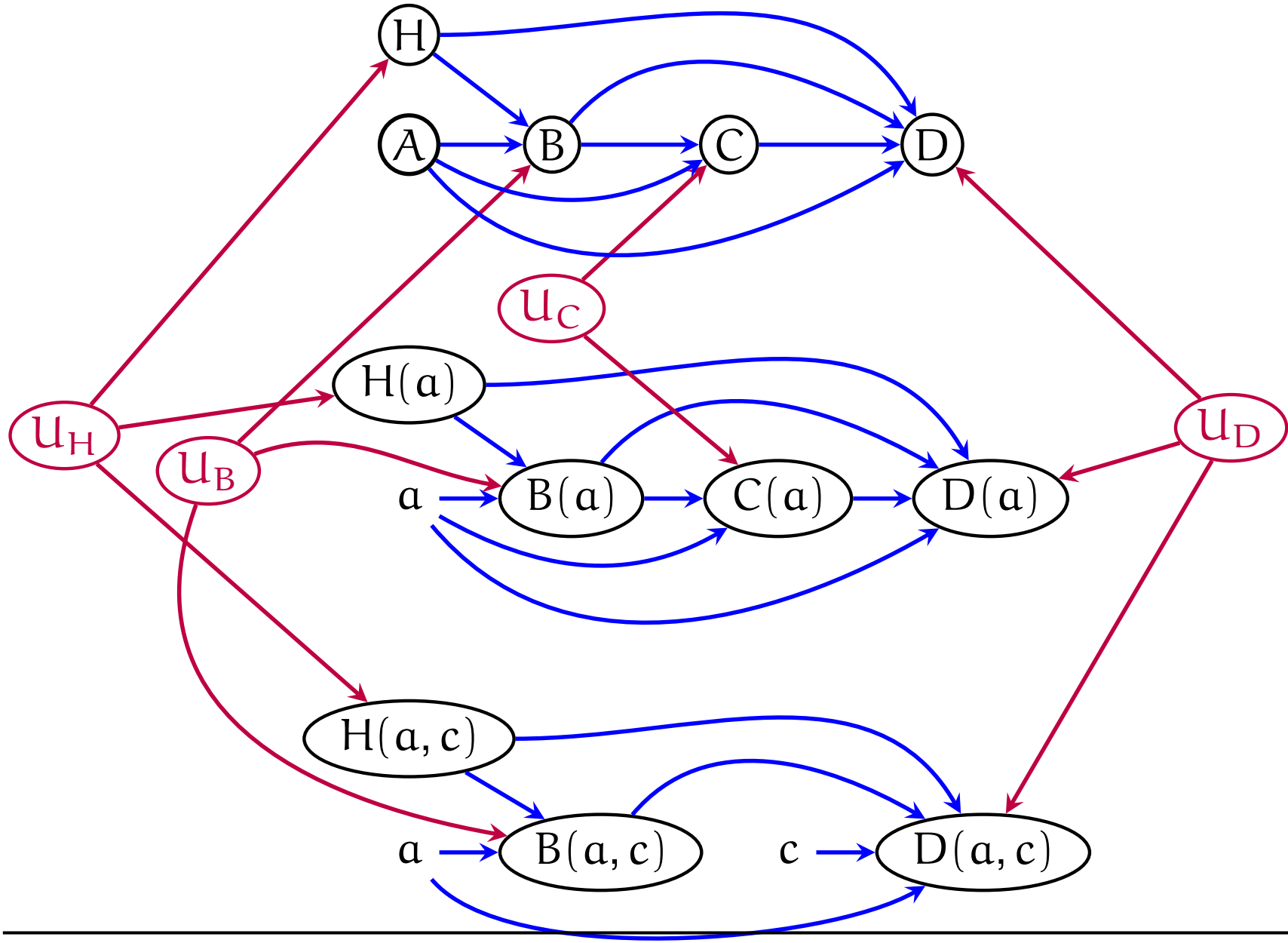
d-separation:

$$A \perp\!\!\!\perp D(\tilde{a}, \tilde{c})$$
$$C(\tilde{a}) \perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \mid B(\tilde{a}), A$$

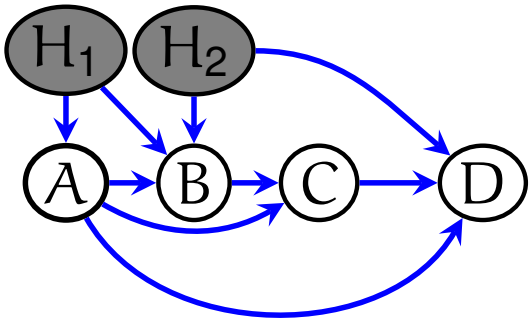
g-formula of Robins (1986) then implies:

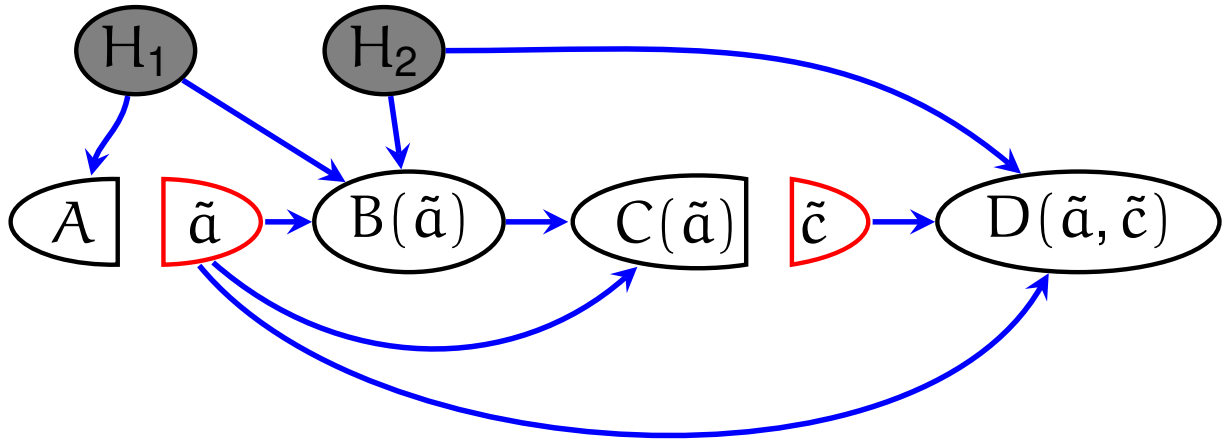
$$P(D(\tilde{a}, \tilde{c}) = d) = \sum_b P(B = b \mid A = \tilde{a}) P(D = d \mid A = \tilde{a}, B = b, C = \tilde{c}).$$

Pearl's Multi-network approach to the same problem



Another example





$$A \perp\!\!\!\perp D(\tilde{a},\tilde{c})$$
$$C(\tilde{a}) \perp\!\!\!\perp D(\tilde{a},\tilde{c}) \mid B(\tilde{a}), A$$

g-formula of Robins (1986) then implies:

$$P(D(\tilde{a},\tilde{c}) = d) = \sum_b P(B = b \mid A = \tilde{a}) P(D = d \mid A = \tilde{a}, B = b, C = \tilde{c}).$$

Can also see that this identification fails if there is a $B \to D$ edge.

Recap: relating Counterfactuals and 'do' notation

Expressions in terms of 'do' can be expressed in terms of counterfactuals:

$$P(Y(x) = y) \equiv P(Y = y \mid do(X = x))$$

Counterfactual notation is more general than 'do' notation.

Ex. Distribution of outcomes that *would* arise among those who took treatment ($X = 1$) had counter-to-fact they not received treatment:

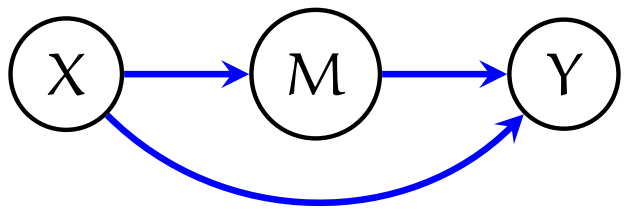
$$P(Y(x = 0) = y \mid X = 1)$$

If treatment is randomized, so $X \perp\!\!\!\perp Y(x = 0)$ then this equals $P(Y(x = 0) = y)$, but in an observational study these may be different.

Relating Counterfactuals and Structural Equations

Potential outcomes can be seen as a different notation for Non-Parametric Structural Equation Models (NPSEMs).

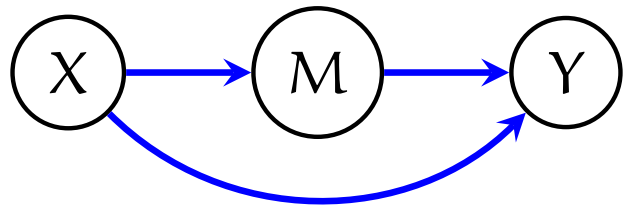
In an NPSEM model associated with a graph each variable is given by an equation expressing the variable as a function of its parents + error term



$$X = f_X(\varepsilon_X)$$
$$M = f_M(X, \varepsilon_M)$$
$$Y = f_Y(X, M, \varepsilon_Y)$$

Relating Counterfactuals and Structural Equations

In an NPSEM model associated with a graph each variable is given by an equation expressing the variable as a function of its parents + error term



But it is clearer to express with potential outcomes

$$\begin{aligned} X &= f_X(\varepsilon_X) \\ M &= f_M(X, \varepsilon_M) \\ Y &= f_Y(X, M, \varepsilon_Y) \end{aligned} \quad \Rightarrow \quad \begin{aligned} X &= f_X(\varepsilon_X) \\ M(x) &= f_M(x, \varepsilon_M) \\ Y(x, m) &= f_Y(x, m, \varepsilon_Y) \end{aligned}$$

observed variables are given by: $M = M(X)$, $Y = Y(X, M(X))$.

Counterfactuals make clear equations represent invariant relationships:

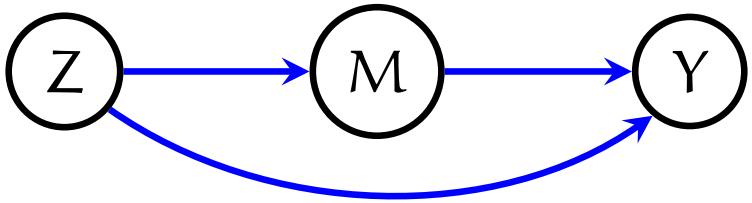
intervening to set X *and* M *to* 0*, the value for* Y *will be:* $f_Y(0, 0, \varepsilon_Y)$.

(Alternative approach via crossing out equations, but this can be confusing since "Y" in the new system is not "Y" in the old system.)

Two important caveats:

- NPSEMs typically assume all variables are seen as being subject to well-defined interventions (not so with potential outcomes)
- Pearl's approach to unifying graphs and counterfactuals simply associates with a DAG the counterfactual model corresponding to an NPSEM with Independent Errors (NPSEM-IEs) aka Structural Causal Models
  Pearl: DAGs and Potential Outcomes are 'equivalent theories'.
- However, any counterfactual independences that can be read from a SWIG will hold under the NPSEM-IE model.
  (Though in general the NPSEM-IE will imply extra independences.)

Pearl's Structural Causal Model / NPSEM-IE approach:



Three binary variable model, no confounding; 7 counterfactual RVs:

$$Z,\ M(z_0), M(z_1),\ Y(m_0, z_0), Y(m_0, z_1), Y(m_1, z_0), Y(m_1, z_1)$$

Pearl's structural causal model assumes cross-world independences:

$$\underbrace{Z}_{\varepsilon_Z} \perp\!\!\!\perp \underbrace{\{M(z_0), M(z_1)\}}_{\varepsilon_M} \perp\!\!\!\perp \underbrace{\{Y(m_0, z_0), Y(m_0, z_1), Y(m_1, z_0), Y(m_1, z_1)\}}_{\varepsilon_Y}$$

Dimension of models:

- No assumptions (allowing confounding): $127 = 2^7 - 1$;
- SWIG (no confounding): 113
- Pearl's NPSEM with indep. errors (no confounding): 19

No. of extra (untestable) counterfactual independence assumptions: 94

How many experimentally untestable assumptions?

Assumption of independent errors implies super-exponentially many 'cross-world' counterfactual independence assumptions:

No. Actual Vars.	2	3	4	K
Dim. $\mathrm{P}(\mathbf{V})$	3	7	15	$2^K - 1$
No. Cnterfactual Vars.	3	7	15	$2^K - 1$
Dim. Cnterfactual Dist.	7	127	32767	$2^{(2^K-1)} - 1$
Dim. SWIG	5	113	32697	$(2^{(2^K-1)} - 1) - \sum_{j=1}^{K-1}(4^j - 2^j)$
Dim. NPSEM-IE	4	19	274	$\sum_{j=0}^{K-1}(2^{2^j} - 1)$
No. untestable indep. constrnts in NPSEM-IE	1	94	32423	$O(2^{2^K-2})$

Table: Dimensions of counterfactual models associated with complete graphs with binary variables.

Critique of Structural Causal Model (SCM) Independent Error Assumption

Pearl's SCM independent error assumption cannot be checked by any randomized experiment on the variables in the graph.

$\Rightarrow$ Connection between experimental interventions and potential outcomes, established by Neyman has been severed;

What about faithfulness and causal discovery procedures?

- Such inferences are explicit that they rely on *faithfulness*;
  - *By contrast: In Pearl's NPSEM-IE approach the simple act of using a DAG is viewed as automatically committing you to making this untestable hypothesis.*
- Predictions (possibly derived assuming faithfulness) regarding *intervention* distributions $\mathrm{P}(\mathrm{Y}(\mathrm{x})) = \mathrm{P}(\mathrm{Y} \,|\, \mathrm{do}(\mathrm{x}))$ *can* be tested by randomized experiments.

Summary so Far

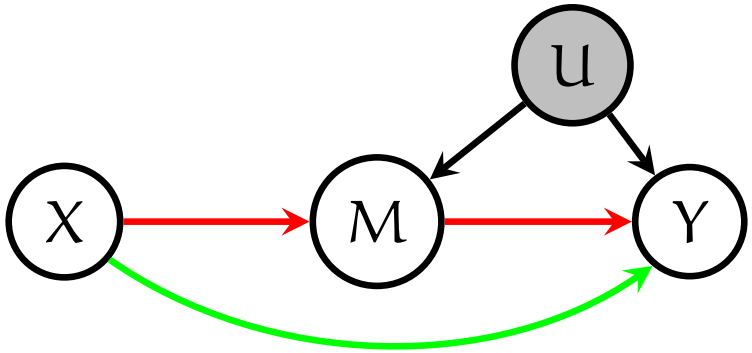
- SWIGs provide a simple way to unify graphs and counterfactuals via node-splitting
- The approach works via linking the factorizations associated with the two graphs.
- The new graph represents a counterfactual distribution that is *identified* from the distribution in the original DAG.
- This provides a language that allows counterfactual and graphical people to communicate.
- (Not covered) Leads to a complete identification algorithm (Extended ID)
  - "Fixing" operation $\Rightarrow$ Splitting $+$ Marginalization
- (Not covered) Can combine information on the absence of individual and population level direct effects.
- (Not covered) Permits formulation of models where interventions on only some variables are well-defined.

Interventional Mediation

Overview of Part Two

- A (fairly) new way to think about mediation and direct effects
  - Separable Direct Effects
- Contrast with other approaches
  - Controlled Direct Effects
  - Principal Stratum Direct Effects
  - Pure (aka Natural) Direct Effects

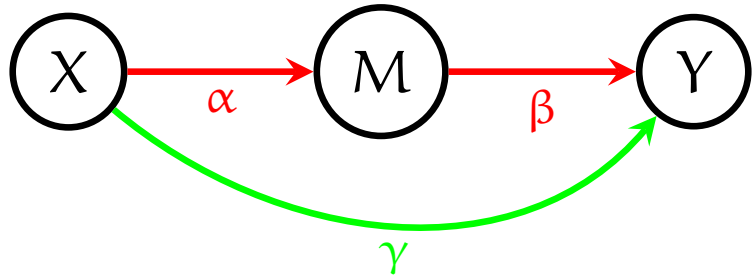
Mediation



Wish to distinguish the 'direct' effect of X on Y from the 'indirect' effect 'via' M.

X is an initial treatment, we assume here randomized;
M is the mediating variable of interest;
Y is the final outcome;
U represents the possibility of unmeasured factors influencing M and Y.

Linear system



$$X = \varepsilon_X$$
$$M = \alpha X + \varepsilon_M$$
$$Y = \beta M + \gamma X + \varepsilon_Y$$

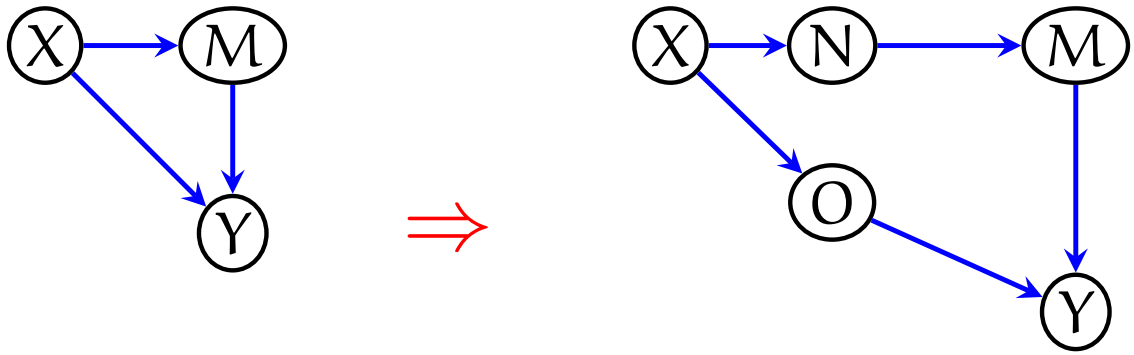
$$Cov(\varepsilon_X, \varepsilon_M) = Cov(\varepsilon_X, \varepsilon_Y) = Cov(\varepsilon_M, \varepsilon_Y) = 0$$

$$E[M(x+1) - M(x)] = \alpha$$
$$E[Y(x+1) - Y(x)] = \underbrace{\gamma}_{\text{direct}} + \underbrace{\alpha\beta}_{\text{indirect}}$$

Qu: How to generalize this story to the non-parametric setting?

Expanded System

All variables binary.

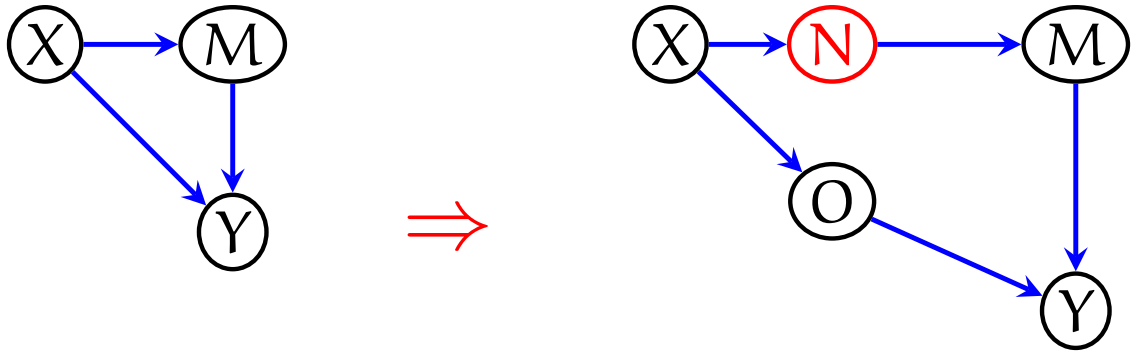


X is Smoking status; M is Hypertension; Y is Myocardial Infarction (MI); Suppose there are additional nodes:

- N = Nicotine exposure;
- O = Other chemical components of cigarettes;

and in addition, the causal graph on the right holds.

Expanded System



Suppose there are additional nodes N, O, with well-defined interventions, such that

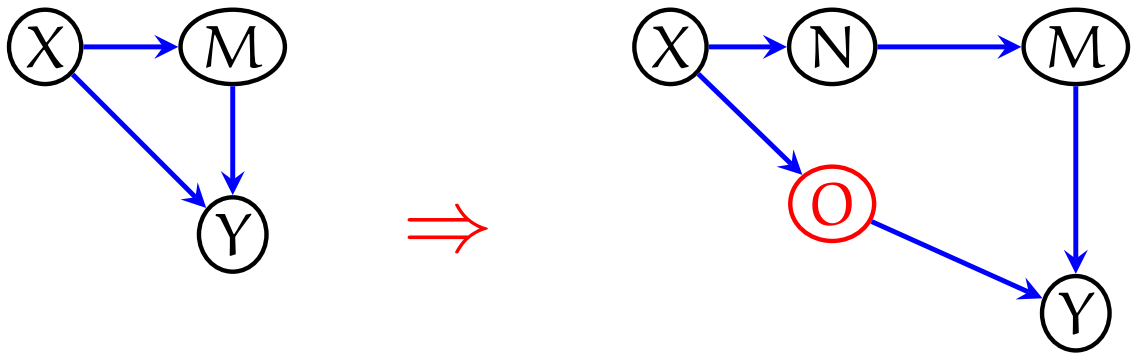
- the effect of X on M is via N;
- the effect of X on Y is via O.

We can then consider contrasts, such as:

$$\text{Direct Effect:} \quad E[Y(x = 1, n = 0) - Y(x = 0, n = 0)]$$

Effect of nicotine-free cigarettes vs. not smoking at all.

Expanded System



Suppose there are additional nodes N, O, with well-defined interventions, such that

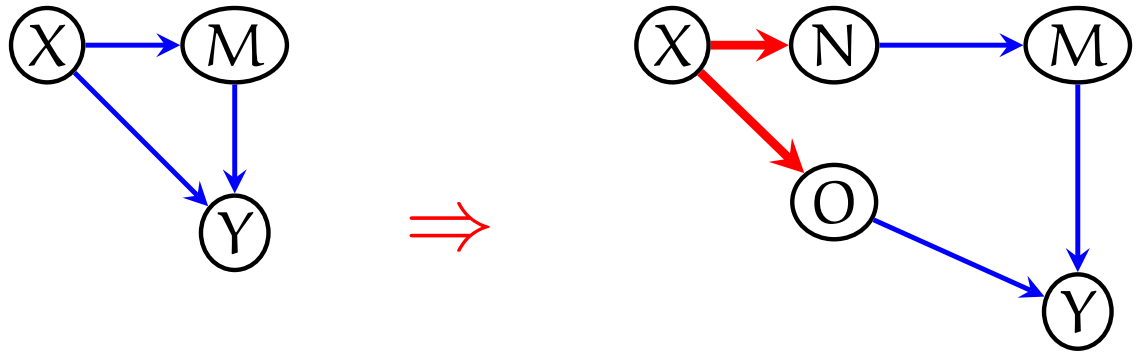
- the effect of X on M is via N;
- the effect of X on Y is via O.

We can then consider contrasts, such as:

$$\text{Direct Effect:} \quad E[Y(x = 1, n = 0) - Y(x = 0, n = 0)]$$
$$\text{Indirect Effect:} \quad E[Y(x = 1, o = 1) - Y(x = 0, o = 1)]$$

Effect of smoking regular cigarettes vs. nicotine-free cigarettes.

Adding Determinism



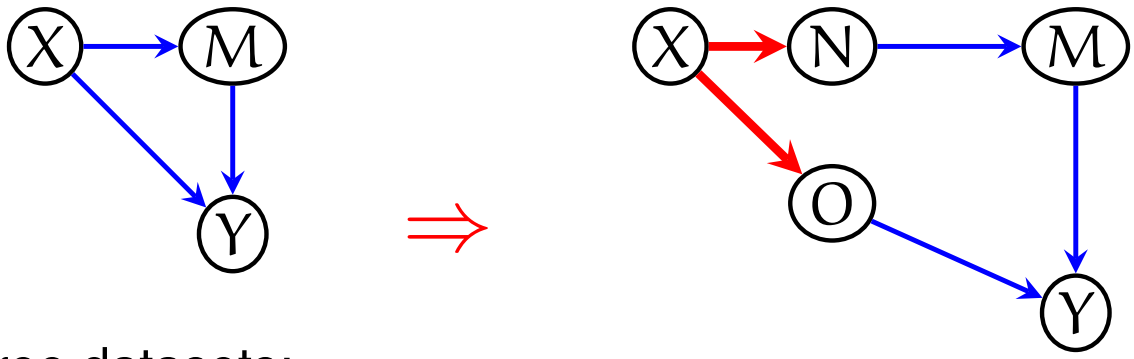
Suppose further that N and O are components of X, so that in the original data $N = X$ and $O = X$, as indicated by the red edges. Then:

$$\begin{aligned} \text{Direct Effect:} \quad & E[Y(x = 1, n = 0) - Y(x = 0, n = 0)] \\ & = E[Y(n = 0, o = 1) - Y(n = 0, o = 0)] \end{aligned}$$

$$\begin{aligned} \text{Indirect Effect:} \quad & E[Y(x = 1, o = 1) - Y(x = 0, o = 1)] \\ & = E[Y(n = 1, o = 1) - Y(n = 0, o = 1)] \end{aligned}$$

Thus $E[Y(x = 1) - Y(x = 0)] =$ Indirect Effect $+$ Direct Effect.
Note that these are contrasts from a four arm (N, O) randomized trial.

Three Datasets



Consider three datasets:

(i) The original observed data from the trial in which X was randomized: X, M, Y, so the data in arm $X = x$ corresponds to $M(x), Y(x)$;

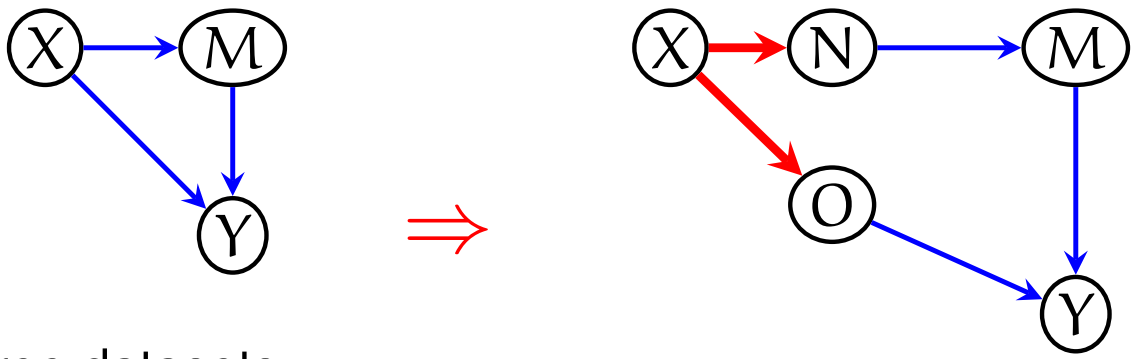
(ii) Data from a putative four arm (N, O) randomized trial; the data in each arm $(n, o) \in \{0, 1\}^2$ corresponds to $M(n, o)$, $Y(n, o)$;

(iii) A dataset obtained from the four arm (N, O) trial (ii) by restricting to the two arms in which $n = o$.

Since in the original data $X = N = O$, (iii) is identified from (i).

Among people with $X = x$ we observe $N = O = x$, hence we observe $M(n = x, o = x)$ and $Y(n = x, o = x)$.This can be viewed as a consistency assumption. As noted by Stensrud this can be tested in a six arm trial.

Three Datasets



Consider three datasets:

(i) The original observed data from the trial in which X was randomized: X, M, Y, so the data in arm $\mathrm{X} = \mathrm{x}$ corresponds to $\mathrm{M}(\mathrm{x}), \mathrm{Y}(\mathrm{x})$;

(ii) Data from a putative four arm (N, O) randomized trial; the data in each arm $(\mathrm{n}, \mathrm{o}) \in \{0, 1\}^2$ corresponds to $\mathrm{M}(\mathrm{n}, \mathrm{o})$, $\mathrm{Y}(\mathrm{n}, \mathrm{o})$;

(iii) A dataset obtained from the four arm (N, O) trial (ii) by restricting to the two arms in which $\mathrm{n} = \mathrm{o}$.

Qu: When is (ii) identified from (i)?

Following Stensrud, we say the effects of N and O on M and Y are *separable*, when this identification holds.

Four arms from two!

Proposition

If the following conditions hold

$$p(Y(n=1,o)=y \mid M(n=1,o)=m) \tag{3}$$
$$= p(Y(n=0,o)=y \mid M(n=0,o)=m) \quad \text{for } o \in \{0,1\};$$

$$p(M(n,o=0)=m) = p(M(n,o=1)=m) \quad \text{for } n \in \{0,1\}; \tag{4}$$

then for $x \in \{0,1\}$ and $\tilde{x} = 1 - x$ we have:

$$p(M(n=x,o=\tilde{x})=m, Y(n=x,o=\tilde{x})=y) \tag{5}$$
$$= p(Y(n=\tilde{x},o=\tilde{x})=y \mid M(n=\tilde{x},o=\tilde{x})=m)p(M(n=x,o=x)=m).$$

Note the last two terms are identified by the data in (iii), the two arm trial in which $n = o$. Thus under the conditions (3) and (4) we can identify (ii) from (iii).

Proof:

$$p(M(n=x,o=\tilde{x}), Y(n=x,o=\tilde{x}))$$
$$= p(Y(n=x,o=\tilde{x}) \mid M(n=x,o=\tilde{x}))p(M(n=x,o=\tilde{x}))$$
$$= p(Y(n=\tilde{x},o=\tilde{x}) \mid M(n=\tilde{x},o=\tilde{x}))p(M(n=x,o=x)).$$

Pearl's Mediation Formula Recovered

Corollary

If the following conditions hold

$$\begin{aligned} p(Y(n=1,o)=y \mid M(n=1,o)=m) \qquad & (1) \\ = p(Y(n=0,o)=y \mid M(n=0,o)=m) \quad & \text{for } o \in \{0,1\}; \end{aligned}$$

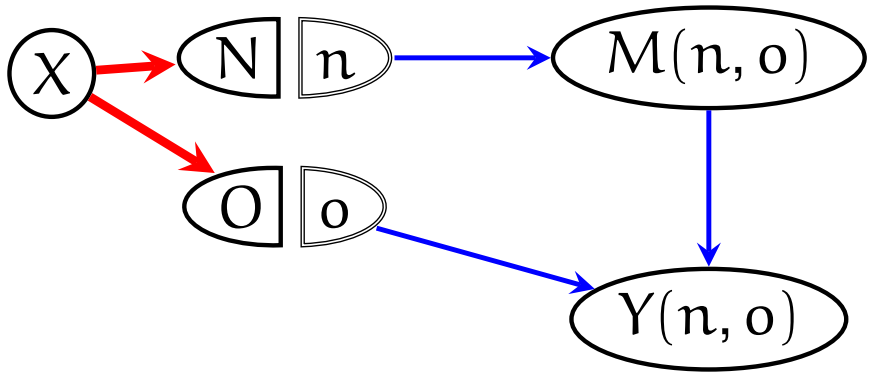
$$p(M(n,o=0)=m) = p(M(n,o=1)=m) \quad \text{for } n \in \{0,1\}; \qquad (2)$$

then for $x \in \{0,1\}$ and $\tilde{x} = 1 - x$ we have:

$$\begin{aligned} & p(Y(n=x,o=\tilde{x})=y) \\ & = \sum_m p(Y(n=\tilde{x},o=\tilde{x})=y \mid M(n=\tilde{x},o=\tilde{x})=m) p(M(n=x,o=x)=m) \\ & = \sum_m p(Y(\tilde{x})=y \mid M(\tilde{x})=m) p(M(x)=m) \\ & = \sum_m p(Y=y \mid M=m, X=\tilde{x}) p(M=m \mid X=x). \qquad (6) \end{aligned}$$

(6) is Pearl's mediation formula proposed to identify pure (aka natural) direct effects.

Assumptions for separability



The assumptions we required for separability are:

$$p(Y(n{=}1, o) = y \,|\, M(n{=}1, o) = m) = p(Y(n{=}0, o) = y \,|\, M(n{=}0, o) = m) \quad \text{for } o \in \{0, 1\};$$

$$p(M(n, o{=}0) = m) = p(M(n, o{=}1) = m) \quad \text{for } n \in \{0, 1\}.$$

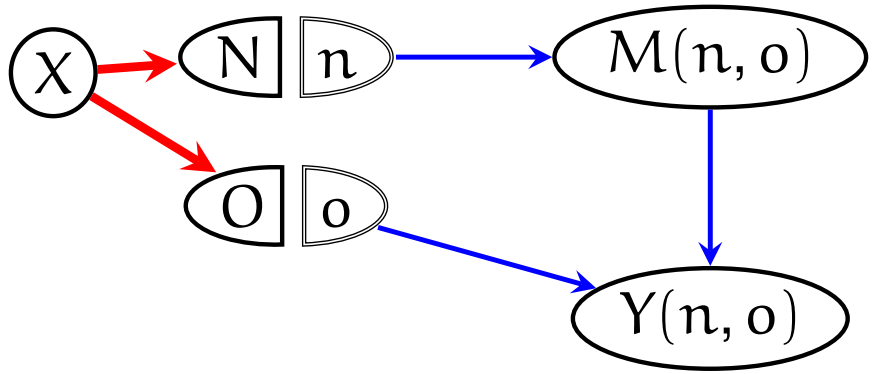
Can be read off the SWIG above with potential outcomes calculus (Malinsky+S+R, 2019).

n is d-separated from $Y(n, o)$ given $M(n, o)$; while o is d-separated from $M(n, o)$.

If there is an unmeasured confounder between M and Y then the first assumption will fail.

These conditions also follow from NPSEM (individual level) no direct effect assumptions.

Testable Assumptions for separability



The assumptions we required for separability are:

$$p(Y(n=1,o)=y \mid M(n=1,o)=m) = p(Y(n=0,o)=y \mid M(n=0,o)=m) \quad \text{for } o \in \{0,1\};$$

$$p(M(n,o=0)=m) = p(M(n,o=1)=m) \quad \text{for } n \in \{0,1\}.$$

Key Observation:
If (in the future), we carry out the four-arm (N, O) study (ii), then these assumptions are subject to direct empirical test.

Contrast with other Approaches to Meidation

The interventional approach to mediation avoids several obstacles present in other approaches:

- Unlike the Pure (or Natural) Direct Effect and Controlled Direct Effect, the interventional approach does **not** require well-defined interventions or counterfactuals on mediators;
- Identification assumptions in the interventional approach are (in principle) subject to empirical test, unlike the (cross-world independence) assumptions used to identify the Pure Direct Effect under an NPSEM with independent errors;
- Unlike the Principal Stratum Direct Effect, the interventional notions of direct and indirect effect are not restricted solely to those subpopulations in which the treatment has no effect on the mediator (at the individual level);

Related Interventional Approaches to Mediation

- Didelez (2019) extends the approach to the context of survival analysis and presents concrete examples of treatment decompositions.
- Martinussen and Stensrud (2020) consider estimation of separable effects when there are competing risks.
- The formulation of a notion of direct effects that do not pass through a mediator *without* the need for well-defined interventions on that mediator also motivates the Organic Mediation approach of Lok (2016, 2020).

Summary

- Given components (N, O) of treatment we can formulate contrasts:

$$\mathrm{E}[\mathrm{Y}(\mathrm{n}=0, \mathrm{o}=1) - \mathrm{Y}(\mathrm{n}=0, \mathrm{o}=0)] \qquad \mathrm{E}[\mathrm{Y}(\mathrm{n}=1, \mathrm{o}=0) - \mathrm{Y}(\mathrm{n}=0, \mathrm{o}=0)];$$

- If N only affects the mediator, and O only affects Y then the contrasts correspond to direct and indirect effects;
- This approach requires that the counterfactuals $\mathrm{M}(\mathrm{n}, \mathrm{o})$ and $\mathrm{Y}(\mathrm{n}, \mathrm{o})$ be well-defined;
  - $\Rightarrow$ requires N and O correspond to substantive variables that could (in principle) be intervened on;
- We gave conditions under which the distribution of $\mathrm{P}(\mathrm{Y}(\mathrm{n} = \mathrm{x}, \mathrm{o} = \tilde{\mathrm{x}}))$ is identified from the data from a randomized experiment;
- The conditions are empirically testable given data from a (subsequent) four-arm trial in which N and O are both randomized.

References for Part One: SWIGs

- Malinsky, D, Shpitser, I, Richardson. TS. A potential outcomes calculus for identifying conditional path-specific effects. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019.
- Pearl, J. Causal diagrams for empirical research, *Biometrika* 82, 4, 669–709, 1995.
- Richardson, TS, Robins, JM. Single World Intervention Graphs. *CSSS Technical Report No. 128* `http://www.csss.washington.edu/Papers/wp128.pdf`, 2013.
- Robins, JM A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling* 7, 1393–1512, 1986.
- Robins, JM, VanderWeele, TJ, Richardson TS. Discussion of "Causal effects in the presence of non compliance a latent variable interpretation" by Forcina, A. *Metron* LXIV (3), 288–298, 2007.
- Shpitser, I, Richardson, TS, Robins, JM. Multivariate Counterfactual Systems and Causal Graphical Models. Arxiv `2008.06017`.
- Spirtes, P, Glymour, C, Scheines R. *Causation, Prediction and Search.* Lecture Notes in Statistics 81, Springer-Verlag, 1993
- LaTeX/Ti*kZ* package on CTAN: `swigs`

References for Part Two: Mediation with SWIGs

- V. Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome *Lifetime Data Analysis*, 25: 593–610, 2019.
- J. J. Lok. Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Statistics in Medicine*, 35, 4008–4020, 2016.
- J. J Lok. Organic direct and indirect effects with post-treatment common causes of mediator and outcome. *arxiv:1510.02753*, 2020.
- J. Pearl. Direct and indirect effects. *UAI*-01, 411–42, 2001.
- J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3: 143–155, 1992.
- J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, and K. Ornstein, editors, *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, chapter 6, pages 1–52. OUP, 2011.
- J. M. Robins, T. S. Richardson, and I. Shpitser. An interventionist approach to mediation analysis. *arxiv:2008.06019*, 2020.
- M. J. Stensrud, J. M. Robins, A. Sarvet, E. J. Tchetgen Tchetgen, and J. G. Young. Conditional separable effects. *arxiv:2006.15681*, 2020.