

# Causal inference and control

Week 6:

Potential outcomes

Søren Wengel Mogensen  
Department of Automatic Control  
soren.wengel\_mogensen@control.lth.se



**LUND**  
UNIVERSITY

## Assignment 2

## Assignment 2

In Problem B, we were considering a *linear Gaussian SCM*. For each  $j = 1, \dots, d$ ,

$$X_j = f_j(\mathbf{PA}_j, N_j) = \sum_{X_i} a_{ji} X_i + N_j$$

where the sum is over  $\mathbf{PA}_j$  and  $N_j$  is Gaussian.

# Valid adjustment sets

## Proposition (ECI, Proposition 6.41)

Assume that  $Y \notin \mathbf{PA}_X$ . The following sets are valid adjustment sets for  $(X, Y)$ .

- $\mathbf{Z} = \mathbf{PA}_X$ .
- $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$  such that  $\mathbf{Z}$  contains no descendants of  $X$  and blocks all back-door paths between  $X$  and  $Y$ .
- no member of  $\mathbf{Z}$  is a descendant of any  $W \in \mathbf{V} \setminus \{X\}$  which lies on a directed path from  $X$  to  $Y$ , and  $\mathbf{Z}$  blocks all nondirected paths between  $X$  and  $Y$ .

We say that  $W$  is a *descendant* of  $X$  if there exists a directed path  $X \rightarrow \dots \rightarrow W$ . We say that a path between  $X$  and  $Y$  is a *back-door path* if  $X \leftarrow \dots Y$ . We say that  $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$  *blocks* a path between  $X$  and  $Y$  if the path is not  $d$ -connecting given  $\mathbf{Z}$ .

# Adjustment in linear structural causal model

Let  $\mathbf{Z}$  be a valid adjustment set for  $(X, Y)$ . In a zero-mean linear Gaussian SCM, we have

$$E(Y|X = x, \mathbf{Z} = \mathbf{z}) = ax + \mathbf{b}^t \mathbf{z}$$

We also have (ECI Problem 6.63)

$$\frac{\partial}{\partial x} E^{do(x)}(Y) = a$$

for a constant  $a$ .

Each valid adjustment  $\mathbf{Z}$  gives us a least-squares estimator,  $\hat{\tau}_{yx}^{\mathbf{z}}$ , of  $a$  using the above regression. So what's the optimal adjustment set in terms of (asymptotic) variance of the estimator?

# Optimal adjustment set

As mentioned there is a solution to this problem (they dispense with the assumption of Gaussian error variables) in

- Henckel, Leonard, Emilija Perković, and Marloes H. Maathuis, *Graphical criteria for efficient total effect estimation via adjustment in causal linear models*. Journal of the Royal Statistical Society Series B: Statistical Methodology 84.2 (2022): 579-599.

We will describe the optimal set defined in this paper (their result also holds for more general graphs and for non-singleton  $X$  and  $Y$ ).

# Optimal adjustment set

We define first the *causal nodes*,  $\text{cn}(X, Y, \mathcal{G})$ , relative to  $(X, Y)$  in the graph  $\mathcal{G}$  as all nodes on a causal (directed) path from  $X$  to  $Y$  except  $X$  itself. We say that  $Z$  is a *descendant* of  $W$  if there exists a directed path from  $W$  to  $Z$  or if  $Z = W$ . We use  $\text{de}(\mathbf{Z})$  to denote the set of descendants of the set  $\mathbf{Z}$ .

We define the *forbidden nodes* relative to  $(X, Y)$  in the graph  $\mathcal{G}$  as

$$\text{forb}(X, Y, \mathcal{G}) = \text{de}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup \{X\}.$$

We define

$$\mathbf{O}(X, Y, \mathcal{G}) = \text{pa}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \setminus \text{forb}(X, Y, \mathcal{G}).$$

# Optimal adjustment set

Theorem (Theorem 3.13, Henckel et al.)

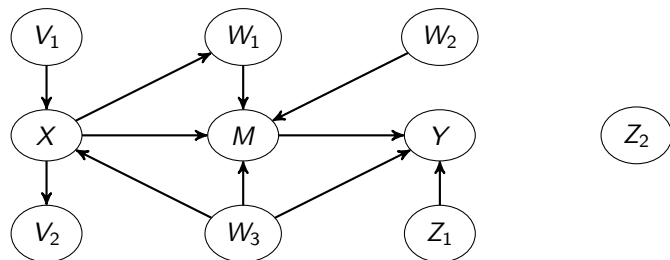
*Assume  $Y$  is a descendant of  $X$  in  $\mathcal{G}$ . The set  $\mathbf{O}(X, Y, \mathcal{G})$  is a valid adjustment set and it attains the minimal asymptotic variance among all valid adjustment sets,  $\mathbf{Z}$ ,*

$$a.\text{var}(\hat{\tau}_{yx}^{\mathbf{O}}) \leq a.\text{var}(\hat{\tau}_{yx}^{\mathbf{Z}}).$$

*Under faithfulness and for a valid adjustment  $\mathbf{Z}$ , if  $a.\text{var}(\hat{\tau}_{yx}^{\mathbf{O}}) = a.\text{var}(\hat{\tau}_{yx}^{\mathbf{Z}})$ , then  $\mathbf{O} \subseteq \mathbf{Z}$ .*



## Optimal adjustment set, example



We first find the causal nodes,  $\text{cn}(X, Y, \mathcal{G}) = \{W_1, M, Y\}$ . We then find  $\text{pa}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\}$ , and  $\text{forb}(X, Y, \mathcal{G}) = \text{de}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup \{X\} = \{W_1, M, Y, X\}$ . Finally,

$$\mathcal{O}(X, Y, \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\} \setminus \{W_1, M, Y, X\} = \{W_2, W_3, Z_1\}.$$

## Graphical marginalization

# Graphical marginalization

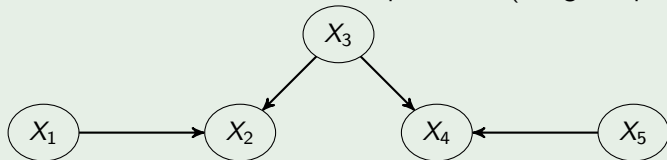
In causal modeling, the idea of *hidden* variables is central. In Week 4, we looked at identification methods that do not require full observation.

In causal discovery, we may be interested in learning a ‘marginal’ of the causal graph when there are hidden variables,  $\mathbf{H}$ , as well as observed variables,  $\mathbf{O}$ . One natural requirement is that the implied conditional independences are the same when restricting to  $\mathbf{O}$ .

# Graphical marginalization

Example (DAGs are not closed under marginalization, Richardson and Spirtes [2002])

Assume  $X_3$  is unobserved. There is no DAG on nodes  $\{X_1, X_2, X_4, X_5\}$  that encodes the same conditional independences (using  $d$ -separation).



# Acyclic directed mixed graphs

We say that a graph is a *directed acyclic mixed graph* (ADMG) if every edge is either *directed*,  $\rightarrow$ , or *bidirected*,  $\leftrightarrow$ .

The extension of *d*-separation to ADMGs is known as *m*-separation.

# Latent projection

Let  $\mathcal{G} = (\mathbf{V}, E)$  be an ADMG,  $\mathbf{V} = \mathbf{O} \cup \mathbf{H}$ . We define the following transformation.

## Definition (Latent projection)

We define  $m(\mathcal{G}, \mathbf{O})$  as the graph such that for  $X, Y \in \mathbf{O}$

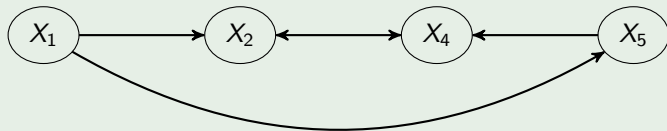
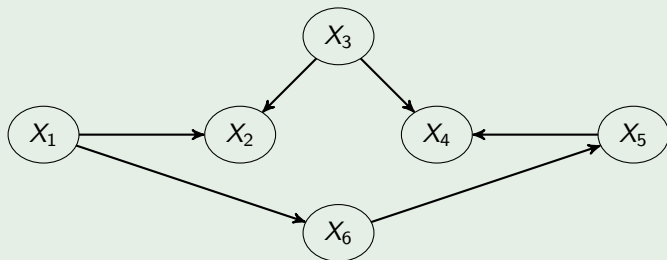
- $X \rightarrow Y$  in  $m(\mathcal{G}, \mathbf{O})$  if there is a directed path  $X \rightarrow \dots \rightarrow Y$  in  $\mathcal{G}$  such that every non-endpoint node is in  $\mathbf{H}$ ,
- $X \leftrightarrow Y$  in  $m(\mathcal{G}, \mathbf{O})$  if there is a path between  $X$  and  $Y$  such that all non-endpoint nodes are in  $\mathbf{H}$ , all non-endpoint nodes are non-colliders, and there are arrowheads at both  $X$  and  $Y$ .

The latent projection is also an ADMG!

# Latent projection

## Example

Let  $\mathbf{V} = \{X_1, X_2, X_3, X_4, X_5, X_6\}$  and  $\mathbf{O} = \{X_1, X_2, X_4, X_5\}$



# Latent projection as a marginal

Let  $\mathcal{G} = (\mathbf{O} \cup \mathbf{H}, E)$  be a DAG and let  $\mathcal{M} = m(\mathcal{G}, \mathbf{O})$ .

## Proposition

*Let  $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{O}$ . We have that  $\mathbf{A}$  and  $\mathbf{B}$  are  $d$ -separated by  $\mathbf{C}$  in  $\mathcal{G}$  if and only if  $\mathbf{A}$  and  $\mathbf{B}$  are  $m$ -separated by  $\mathbf{C}$  in  $m(\mathcal{G}, \mathbf{O})$ .*



Potential outcomes

# SCM, 'procedural' explanation

A central part of the definition of a structural causal model is the *modularity* of assignment mechanisms,  $f_j$ , i.e., they remain the same unless we intervene upon them.

$$X_1 = f_1(PA_1, N_1) = \bar{f}_1(N_1)$$

$$X_2 = f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2)$$

$$X_3 = f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3)$$

...

$$X_{d-1} = f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1})$$

$$X_d = f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)$$

# Interventions, 'procedural' explanation

If we have an (atomic) intervention  $do(X_j = a)$ ,

$$X_1 = f_1(PA_1, N_1) = \bar{f}_1(N_1)$$

$$X_2 = f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2)$$

$$X_3 = f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3)$$

...

$$X_j = a$$

...

$$X_{d-1} = f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1})$$

$$X_d = f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)$$

# Potential outcomes

The *potential outcomes* framework (or *Rubin causal model*) takes a different approach to defining causal notions [Neyman, 1923, Fisher, 1925, Rubin, 1974].

## ECI Example 3.4

We start by looking at Example 3.4. We assume we have a population of patients with a specific eye disease. There is a treatment,  $T$ , and 99% of patients are cured ( $B = 0$ ) if treated ( $T = 1$ ), and otherwise they go blind ( $B = 1$ ). The 1% go blind if and only if they are treated. The difference in effect of treatment is explained by a rare condition,  $N_B$ . Treatment is independent of  $N_B$  as this is unknown to the doctor.

$$T = N_T$$

$$B = T \cdot N_B + (1 - T) \cdot (1 - N_B)$$

and  $N_B \sim \text{Ber}(0.01)$ .

# Potential outcomes

The potential outcomes framework uses a different formalism to specify a causal model. Assume we have a fixed population of individuals (or *units*),  $n = 200$ . Each of these individuals,  $u$ , have *two* potential outcomes, one for each level of treatment.

$B_u(t = 0)$  is the value of  $B$  unit  $u$  would have had had they received treatment.  $B_u(t = 1)$  is the value of  $B$  unit  $u$  would have had had they not received treatment. These are deterministic (could also be more general random variables). One of these are *counterfactual* in the sense that it is unobservable and can actually never be observed.

SUTVA (stable-unit treatment value assumption) is commonly made.

# Potential outcomes

$u$	$T$	$B_u(t=0)$	$B_u(t=1)$	$B_u(t=1) - B_u(t=0)$
1	1	1	0	-1
2	1	1	0	-1
3	0	0	1	1
$\vdots$				

Grey is observed information. The quantity

$$B_u(t=1) - B_u(t=0)$$

is the *unit-level causal effect*. The *average causal effect* is

$$\frac{1}{n} \sum_{u=1}^n B_u(t=1) - B_u(t=0)$$

# Potential outcomes

Assume we have data from a randomized experiment. If  $u \in U_0$  received no treatment and  $u \in U_1$  received treatment then

$$\frac{1}{|U_1|} \sum_{u \in U_1} B_u(t=1) - \frac{1}{|U_0|} \sum_{u \in U_0} B_u(t=0)$$

is an unbiased estimator of the average causal effect.



# SUTVA

Assume that  $\mathbf{T} = (T_1, \dots, T_n)$  is a vector of treatment assignments. In our example, SUTVA (stable-unit treatment value assumption) is the assumption that

$$B_u(\mathbf{T}) = B_u(\mathbf{T}') \quad \text{if} \quad T_u = T'_u.$$

for all  $\mathbf{T}, \mathbf{T}'$  and that for all  $\omega \in \Omega$

$$T_u(\omega) = t \Rightarrow B_u(T = t)(\omega) = B_u(\omega)$$

(if unit  $u$  receives treatment  $t$ , then the observed outcome is  $B_u(T = t)$ ). This first is 'no interference' and the second is known as consistency (the table implicitly used these assumptions).

# Ignorability

In this framework, the treatment assignment *unconfounded* or *ignorable* if

$$B(T = 0), B(T = 1) \perp T$$

(we lost the unit subscripts).

There is also a conditional version of this assumption. Under this (and the previous assumptions),

$$E(B|T = t) = E(B(T = t)|T = t) = E(B(T = t))$$

so the average treatment effect is identified,

$$E(B(T = 1) - B(T = 0)) = E(B|T = 1) - E(B|T = 0).$$

# References I

- Ronald Aylmer Fisher. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, UK, 1925.
- Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9 (translated). *Statistical Science*, 5: 465–480, 1923.
- Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688, 1974.