Causal inference and control

Week 1:

Structural causal models

Søren Wengel Mogensen
Department of Automatic Control
soren.wengel_mogensen@control.lth.se

LUND
UNIVERSITY

Course overview

- Lectures Mondays 10.15-12.00, exercise classes Wednesday 10.15-12.00 (no class in Easter week and on public holidays).
- We will use the book Elements of Causal Inference (free electronic version).
- Course webpage https://github.com/soerenwengel/causalinferenceandcontrol/wiki/Overview

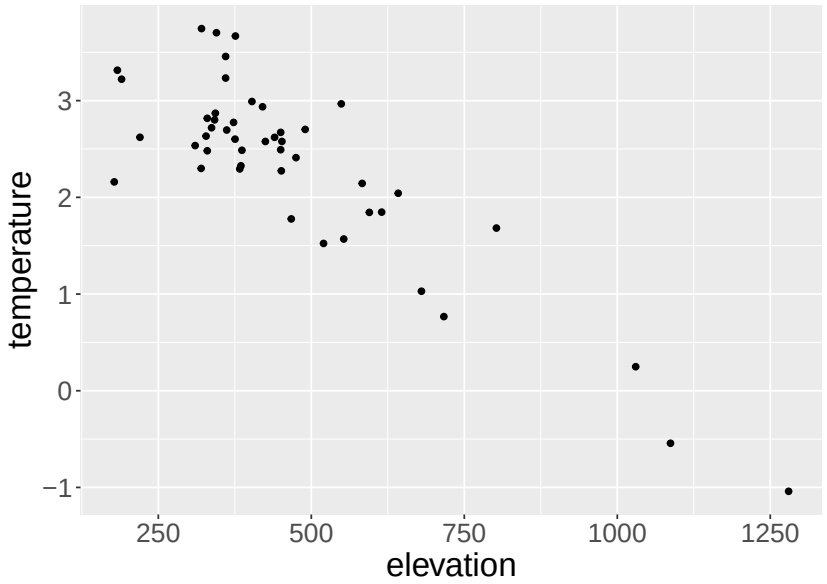
Predictive methods

Many methods from statistics, machine learning, and other quantitative sciences aim to predict a target variable, Y, from a set of covariates/features, X. This uses samples from the joint distribution of (X, Y) to learn how to map observed covariates to an educated guess of an unseen Y.

At training time, we have $(X_1, Y_1,), \ldots, (X_n, Y_n)$, at test time we only have $\bar{X}_1, \ldots, \bar{X}_m$ and the task is to learn f such that $f(\bar{X}_i)$ and $\bar{Y}_i$ are close in some sense.

The target/feature divide is dictated by the problem at hand and by what data is available when the prediction is needed (for instance, this is the case for weather forecasting, spam filtering, early warning systems, and diagnostic models).

Example



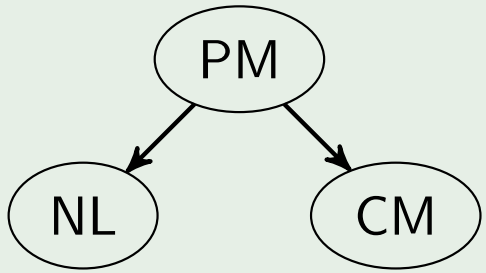
This is mean annual temperature, T, and elevation, E, of weather stations in Jämtland (Swedish province). We can use E to predict T, and the other way around. What would happen if we were to magically intervene on E, or on T, and predict after the intervention?

Structural causal model, example

Structural causal models are one way to formalize causal notions and we will define them shortly. First, an example with three 'endogenous' variables (PM: parent myopia, NL: night light, CM: child myopia) and three noise variables (N_{PM}, N_{NL}, N_{CM}).

Example (Myopia)

$$PM = N_{PM}$$
$$NL = f(PM, N_{NL})$$
$$CM = g(PM, N_{CM})$$



Structural causal model

Definition (Structural causal model)

We consider a set of 'endogenous' variables $\{X_1, \ldots, X_d\}$ and a set of independent noise variables $\{N_1, \ldots, N_d\}$. For each $j = 1, \ldots, d$, $PA_j \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$. A *structural causal model*, $\mathcal{C} = (S, P_N)$, consists of a distribution P_N over the noise variables and a set, S, of acyclic structural assignments,

$$X_j = f_j(PA_j, N_j), j = 1, \ldots, d.$$

Structural causal model

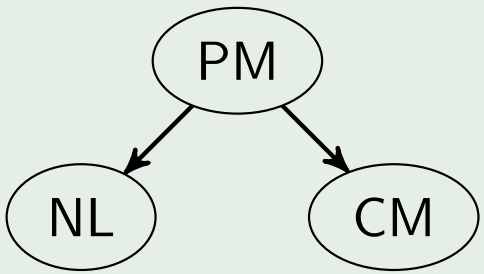
Definition (Structural causal model)

We consider a set of 'endogenous' variables $\{X_1, \ldots, X_d\}$ and a set of independent noise variables $\{N_1, \ldots, N_d\}$. For each $j = 1, \ldots, d$, $PA_j \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$. A *structural causal model*, $\mathcal{C} = (S, P_N)$, consists of a distribution P_N over the noise variables and a set, S, of acyclic structural assignments,

$$X_j = f_j(PA_j, N_j), j = 1, \ldots, d.$$

Example

$$PM = N_{PM}$$
$$NL = f(PM, N_{NL})$$
$$CM = g(PM, N_{CM})$$

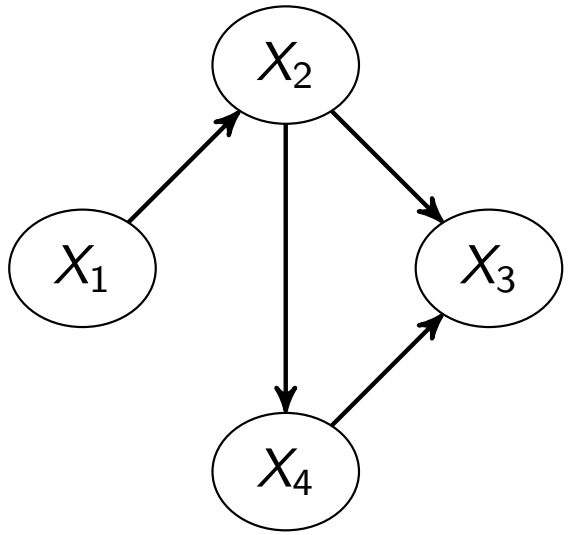


Graph of a structural causal model

We will take an SCM as the fundamental object representing a causal model and from an SCM we define a graph as above.

A *graph*, $\mathcal{G}$, is formally a pair, (V, E), where V is a set of *nodes* (or, *vertices*) ($V = \{X_1, \ldots, X_d\}$ above) and E is a set of edges (the above edges are *directed*, $\rightarrow$) represented by ordered pairs of nodes, (X, Y), such that $X, Y \in V$ and $X \neq Y$. A *path between* X_1 *and* X_{m+1} is a alternating sequence of nodes and edges $(X_1, e_1, X_2, e_2, \ldots, e_m, X_{m+1})$ such that each node occurs at most once.

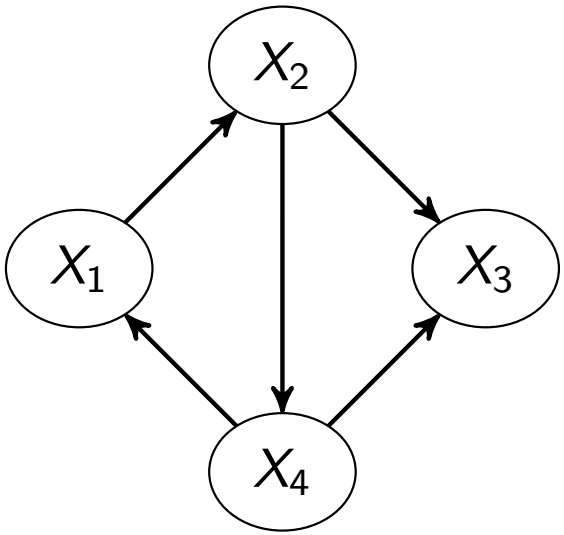
$X_4 \rightarrow X_3 \leftarrow X_2 \leftarrow X_1$ is a path.

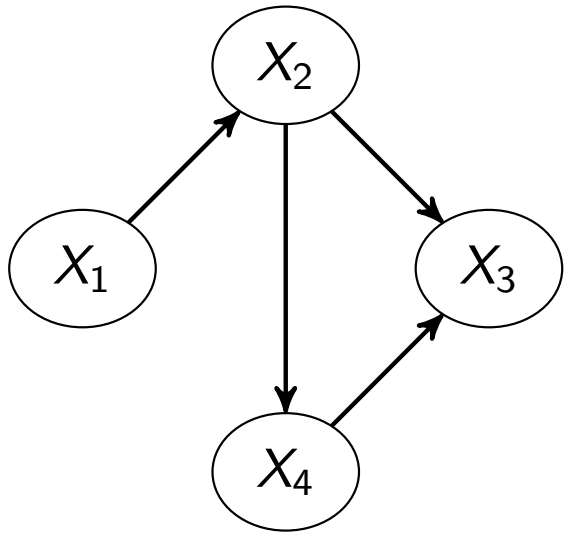


Graph of a structural causal model

A graph is said to be *directed* if it contains only directed edges, $\rightarrow$, and acyclic if there is no *directed path* $X_{i_1} \rightarrow X_{i_2} \rightarrow \ldots \rightarrow X_{i_m} \rightarrow X_{i_1}$ for any i_1. A directed and acyclic graphs is known as a DAG.

If $X_1 \rightarrow X_2$, then we say that X_1 is a *parent* of X_2 and that X_2 is a *child* of X_1.





Graph of a structural causal model

We define a graph, (V, E), from an SCM with $V = \{X_1, X_2, \ldots, X_d\}$. For each j, we include an edge $X_i \to X_j$ for each $X_i \in PA_j$.

We assume throughout that the graph of an SCM is a DAG.

Graph of a structural causal model

We define a graph, (V, E), from an SCM with $V = \{X_1, X_2, \ldots, X_d\}$. For each j, we include an edge $X_i \to X_j$ for each $X_i \in PA_j$.

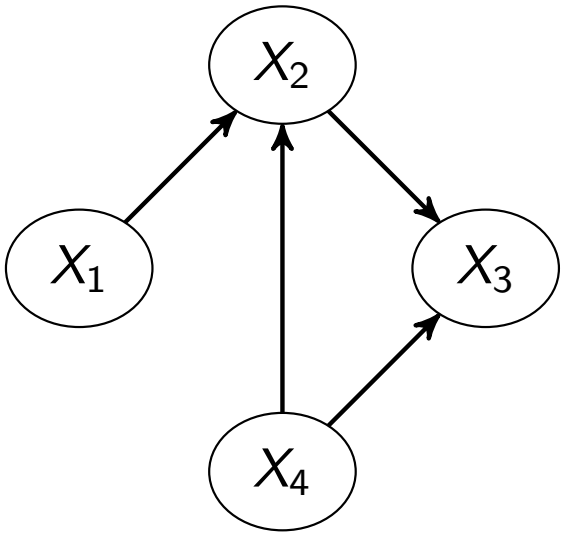
We assume throughout that the graph of an SCM is a DAG.

Recall that

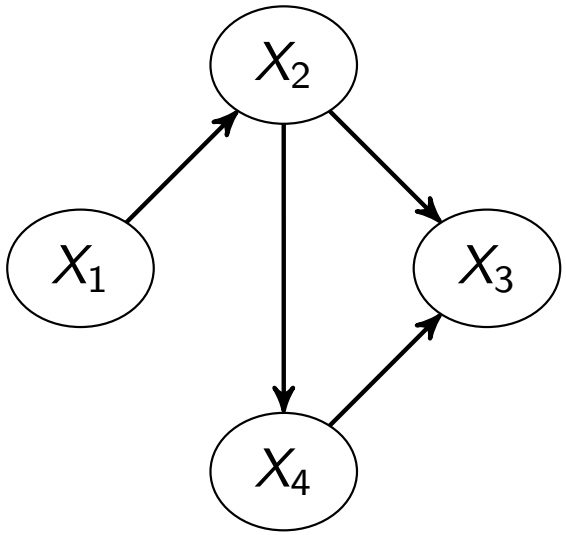
$$X_j = f_j(PA_j, N_j).$$

Graph of a structural causal model

Every DAG as at least one *topological ordering*. This is a numbering, σ, of its vertices such that for every edge $X \to Y$ it holds that $\sigma(X) < \sigma(Y)$.



Node	σ_1	σ_2
X_1	1	2
X_2	3	3
X_3	4	4
X_4	2	1



Node	σ
X_1	1
X_2	2
X_3	4
X_4	3

Interventions

Definition (Interventional distribution)

Let $C = (S, P_N)$ be SCM and let $\tilde{S}$ be a set of structural assignments indexed by $I \subseteq \{1, \ldots, d\}$,

$$X_k = \tilde{f}_k(\tilde{PA}_k, \tilde{N}_k), k \in I.$$

We define an SCM by using the corresponding structural assignment in $\tilde{S}$ if $j \in I$, and otherwise the structural assignment in S, and require that the resulting set of assignments is acyclic. This gives a new SCM, and we say that its distribution is the *interventional* distribution defined by $\tilde{S}$, denoted by $P^{do(X_k=\tilde{f}_k(\tilde{PA}_k,\tilde{N}_k),k\in I)} = P^{do(X_k)}$. We assume that the corresponding density exists and denote this by $p^{do(X_k)}$.

The book uses the notation $P^{C;do(X_k=\tilde{f}_k(\tilde{PA}_k,\tilde{N}_k),k\in I)}$, but we omit the causal model in the notation. The *observational distribution*, $\mathrm{P_X}$, corresponds to $S = \emptyset$ (no intervention). The noise terms are jointly independent, also in the interventional SCM.

Interventions

We say that an intervention, $\tilde{S}$, is *atomic* if $\tilde{S} = \{X_k = a\}$ for $a \in \mathbb{R}$, and write $P_X^{do(X_k=a)}$ for its interventional distribution. In this case, we can let $\tilde{PA}_k = \emptyset$.

SCM, 'procedural' explanation

We assume that the structural causal models are acyclic. This means that there is an *order* of the variables such that for each j and $k > j$, X_k is not a parent of X_j.

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

Note that $\bar{f}_j$ is allowed to depend on all the variables $\{X_1, \dots, X_{j-1}\}$, but may only depend on a subset.

SCM, 'procedural' explanation

We assume that the structural causal models are acyclic. This means that there is an *order* of the variables such that for each j and $k > j$, X_k is not a parent of X_j.

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

We have

$$
\begin{aligned}
p(x_1, \dots, x_n) &= p(x_j \mid x_{j-1}, \dots, x_1) p(x_{j-1} \mid x_{j-2}, \dots, x_1) \dots p(x_2 \mid x_1) x_1 \\
&= \prod_j p(x_j \mid PA_j)
\end{aligned}
$$

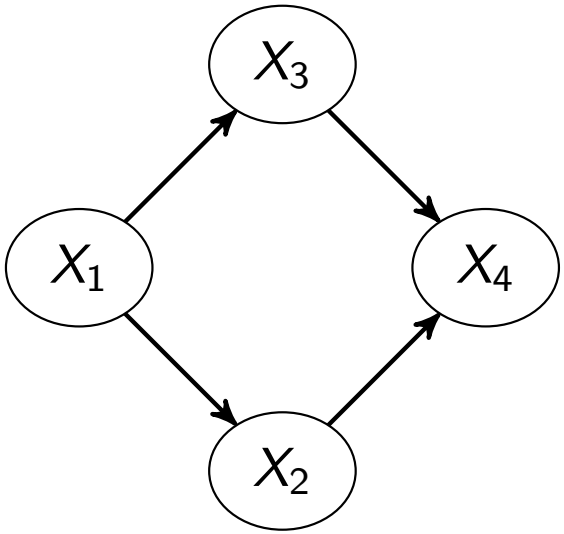
Interventions, 'procedural' explanation

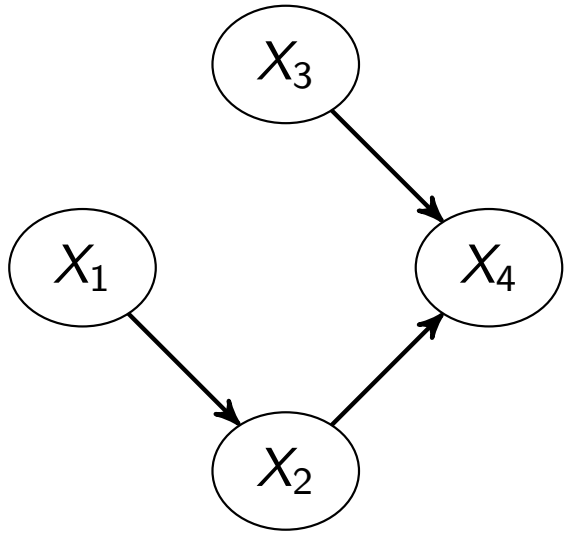
If we have an (atomic) intervention $do(X_j) = a$, we obtain,

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_j &= a \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

Interventions, graph

An interventional SCM corresponding to an atomic intervention, $do(X_k = a)$, is represented by a graph where all edges 'into' X_j are removed.





$do(X_3 = \tilde{N}_3)$

Intervention, example

Example

In the example with elevation, E, and temperature, T, which structural causal model seems to correspond with the physical intuition?

$$E = N_E \qquad T = N_T$$
$$T = f_T(E, N_T) \qquad E = f_E(T, N_E)$$

Intervention, example

Example

In the example with elevation, E, and temperature, T, which structural causal model seems to correspond with the physical intuition?

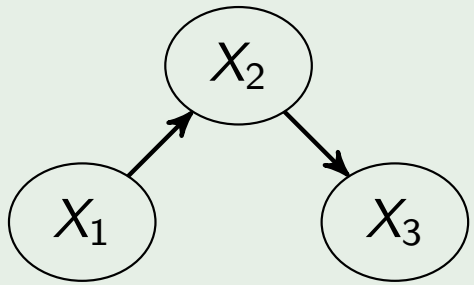
$$E = N_E \qquad T = N_T$$
$$T = f_T(E, N_T) \qquad E = f_E(T, N_E)$$

One should note that the observational distribution may be written in either way and an SCM is not simply a way to specify a distribution.

Intervention, example

Example

$$X_1 = N_1$$
$$X_2 = 2X_1 + N_2$$
$$X_3 = -X_2 + N_3$$



Let $N_i \sim N(0, 1)$. What are $P_X^{do(X_2=1)}$ and $P_X^{do(X_3=-1)}$? How does this compare with conditioning?

Interventions

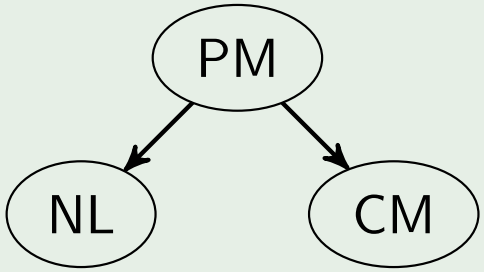
An important observations is that most classical statistical or machine learning models describe a single distribution, the observational distribution. An SCM describes an observational distribution as well as a set of interventional distributions (or formally, they are described by a collection of SCMs).

Example (continued)

Let us say we do not observe PM, or choose to not model it. There is a correlation between NL and CM. Does this mean that we should encourage people to remove night light? We can compare $P^{do(NL=1)}$ and $P^{do(NL=0)}$.

Example (Myopia)

$$PM = N_{PM}$$
$$NL = f(PM, N_{NL})$$
$$CM = g(PM, N_{CM})$$



Truncated factorization

The interventional distributions can be rewritten in terms of the observational distribution. First, we note that

$$p^{do(X_k)}(x_j \mid x_{pa(j)}) = p(x_j \mid x_{pa(j)}), \quad j \neq k.$$

The observational distribution factorizes according to the DAG of the SCM,

$$p(x_1, \ldots, x_d) = \prod_j p(x_j \mid x_{pa(j)}).$$

For $do(X_k = \tilde{N}_k)$, the interventional distribution factorizes as

$$p^{do(X_k)}(x_1, \ldots, x_d) = p^{do(X_k)}(x_k) \prod_{j \neq k} p^{do(X_k)}(x_j \mid x_{pa(j)}) = p^{do(X_k)}(x_k) \prod_{j \neq k} p(x_j \mid x_{pa(j)})$$

This rewrites the interventional distributions as observational distributions.

Causal effects

Using an SCM, we may formalize what is meant by the existence of a *total causal effect*.

Definition (Existence of total causal effect)

We say that there is a *total causal effect* from X to Y if there exists a random variable $\tilde{N}_X$ such that

$$X \not\perp\!\!\!\perp Y \text{ in } P^{do(X=\tilde{N}_X)}.$$

Causal effects

Using an SCM, we may formalize what is meant by the existence of a *total causal effect*.

Definition (Existence of total causal effect)

We say that there is a *total causal effect* from X to Y if there exists a random variable $\tilde{N}_X$ such that

$$X \not\perp\!\!\!\perp Y \text{ in } P^{do(X=\tilde{N}_X)}.$$

Intervene on T,

$$E = N_E$$
$$T = f_T(E, N_T)$$

Intervene on A,

$$E = N_E$$
$$T = f_T(E, N_T)$$

Causal effects

Using an SCM, we may formalize what is meant by the existence of a *total causal effect*.

Definition (Existence of total causal effect)

We say that there is a *total causal effect* from X to Y if there exists a random variable $\tilde{N}_X$ such that

$$X \not\perp\!\!\!\perp Y \text{ in } P^{do(X=\tilde{N}_X)}.$$

Intervene on T,

$$E = N_E$$
$$T = \tilde{N}_T$$

Intervene on A,

$$E = N_E$$
$$T = f_T(E, N_T)$$

Causal effects

Using an SCM, we may formalize what is meant by the existence of a *total causal effect*.

Definition (Existence of total causal effect)

We say that there is a *total causal effect* from X to Y if there exists a random variable $\tilde{N}_X$ such that

$$X \not\perp\!\!\!\perp Y \text{ in } P^{do(X=\tilde{N}_X)}.$$

Intervene on T,

$$E = N_E$$
$$T = f_T(E, N_T)$$

Intervene on A,

$$E = \tilde{N}_E$$
$$T = f_T(E, N_T)$$

Causal effects

Proposition (Existence of total causal effect, Proposition 6.14)

- *If there is no directed path from X to Y, then there is no total causal effect from X to Y.*
- *There may be a directed path, but no total causal effect.*

Causal effects

Proposition (Existence of total causal effect, Proposition 6.14)

- *If there is no directed path from X to Y, then there is no total causal effect from X to Y.*
- *There may be a directed path, but no total causal effect.*

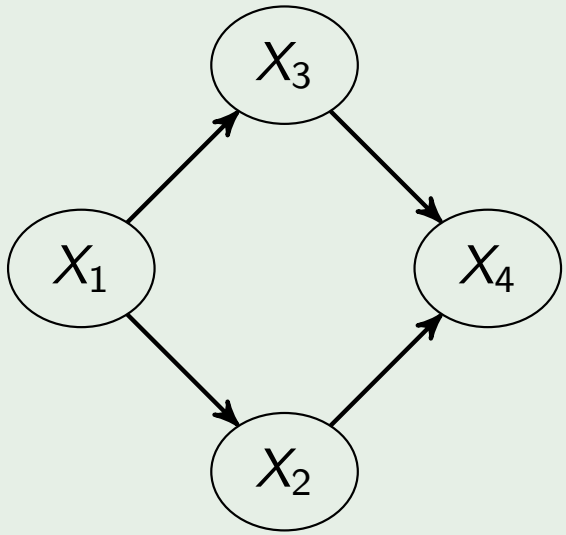
Intuitively, the intervention makes X 'exogenous' and X and Y will be independent in the interventional distribution if there is no sequence $X = X_1, X_2, \ldots, X_m = Y$ such that $X_j = f_j(PA_j, N_j)$ and f_j depends on X_{j-1}.

Causal effects

The following example shows that there may be a directed path from X (X_1) to Y (X_4), but no total causal effect.

Example

$$X_1 = N_1$$
$$X_2 = X_1 + N_2$$
$$X_3 = -X_1 + N_3$$
$$X_4 = X_2 + X_3 + N_4$$

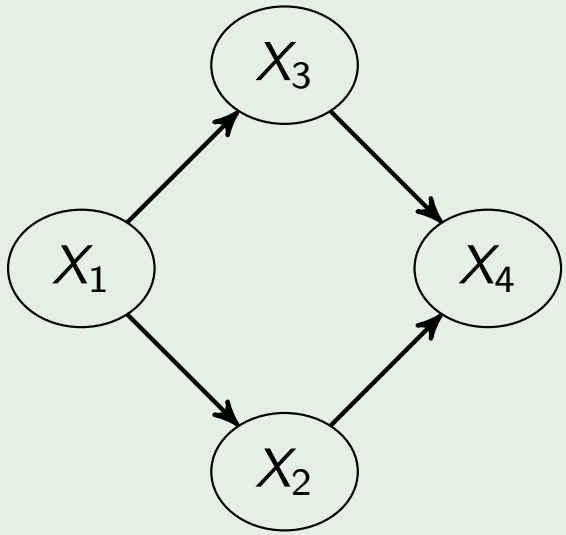


Causal effects

The following example shows that there may be a directed path from X (X_1) to Y (X_4), but no total causal effect.

Example

$$X_1 = N_1$$
$$X_2 = X_1 + N_2$$
$$X_3 = -X_1 + N_3$$
$$X_4 = X_2 + X_3 + N_4 = N_2 + N_3 + N_4$$



Average treatment effect

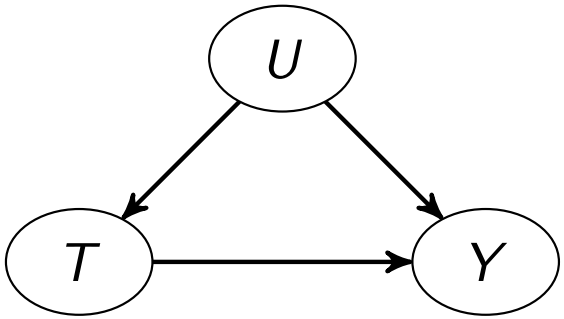
We may of course also be interested in quantifying the size of a causal effect of X, not only deciding its existence. This is done using, e.g., $P^{do(X=a)}$ for different a. Let $E_{do(X=a)}(Y)$ denote the expectation of Y in the interventional distribution $P^{do(X=a)}$. If X is binary

$$E_{do(X=1)}(Y) - E_{do(X=0)}(Y)$$

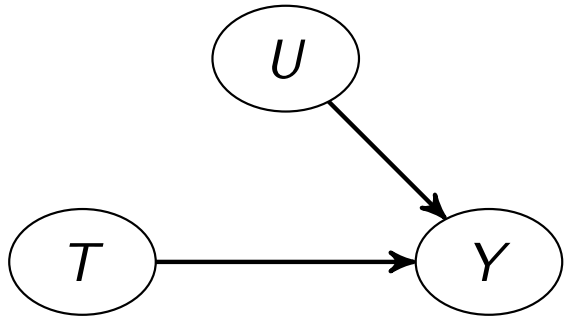
is known as the *average treatment effect*.

Randomized controlled trials

Say we have a *treatment*, T, an *outcome*, Y, and a set of patient characteristics, U. In ordinary medical practice, patient characteristics inform the treatment. Randomized controlled trials intervene on treatment, making it exogenous.



No intervention



Randomized T

This allows consistent estimation of $E_{do(X=1)}(Y) - E_{do(X=0)}(Y)$, even if U is not (fully) observed.

Conditional independence

The graph of an SCM encodes a lot of information, including induced *conditional independences*. Two random variables X_A and X_B are *independent* if

$$p(x_A, x_B) = p(x_A)p(x_B)$$

for all x_A and x_B and they are *conditionally independent* given X_C if

$$p(x_A, x_B \mid x_C) = p(x_A \mid x_C)p(x_B \mid x_C)$$

for all x_A, x_B, x_C such that $p(x_C) > 0$ in which case we write $A \perp\!\!\!\perp B \mid C$.

d-separation

The connection between graphs and conditional independences can be made through the notion of *d-separation*. We say that X is an *ancestor* of Y if there exists a directed path from X to Y and we let $an_{\mathcal{G}}(Y)$ denote the set of ancestors of Y in $\mathcal{G}$. For $C \subseteq V$, we let $an_{\mathcal{G}}(C) = \cup_{Y \in C} an_{\mathcal{G}}(Y)$. A (nonendpoint) node, Z, on a path is a *collider* if both adjacent edges on the path has *heads* at Z $(\ldots \rightarrow Z \leftarrow \ldots)$, and otherwise it is a noncollider.

Definition (*d*-connecting path)

Let $X, Y \in V$ and $C \subseteq V \setminus \{X, Y\}$. We say that a path is *d-connecting* between X and Y given C if every collider is in $an_{\mathcal{G}}(C)$ and no noncollider is in C.

d-separation

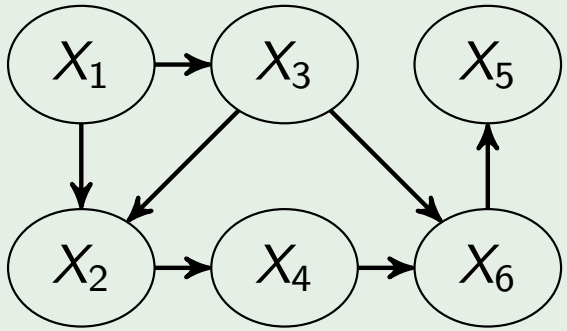
We say that X is an *ancestor* of Y if there exists a directed path from X to Y and we let $an_{\mathcal{G}}(Y)$ denote the set of ancestors of Y in $\mathcal{G}$. For $C \subseteq V$, we let $an_{\mathcal{G}}(C) = \cup_{Y \in C} an_{\mathcal{G}}(Y)$. A (nonendpoint) node, Z, on a path is a *collider* if both adjacent edges on the path as *heads* at Z ($\ldots \to Z \leftarrow \ldots$), and otherwise it is a noncollider.

Definition (d-connecting path)

Let $X, Y \in V$ and $C \subseteq V \setminus \{X, Y\}$. We say that a path is *d-connecting* between X and Y given C if every collider is in $an_{\mathcal{G}}(C)$ and no noncollider is in C.

Example

The path $X_3 \to X_6 \leftarrow X_4 \leftarrow X_2$ is not d-connecting between X_3 and X_2 given $\emptyset$. It is also not d-connecting given $\{X_4, X_6\}$. It is d-connecting given $\{X_5\}$.



d-separation

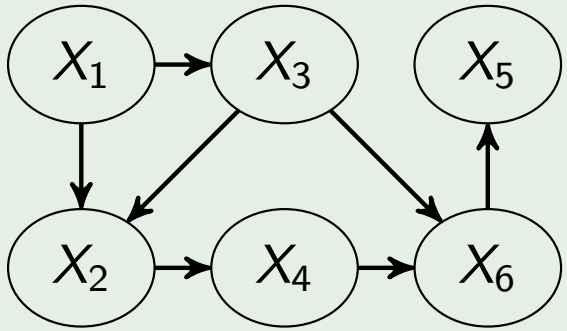
We say that X is an *ancestor* of Y if there exists a directed path from X to Y and we let $an_{\mathcal{G}}(Y)$ denote the set of ancestors of Y in $\mathcal{G}$. For $C \subseteq V$, we let $an_{\mathcal{G}}(C) = \cup_{Y \in C} an_{\mathcal{G}}(Y)$. A (nonendpoint) node, Z, on a path is a *collider* if both adjacent edges on the path as *heads* at Z ($\ldots \to Z \leftarrow \ldots$), and otherwise it is a noncollider.

Definition (d-connecting path)

Let $X, Y \in V$ and $C \subseteq V \setminus \{X, Y\}$. We say that a path is *d-connecting* between X and Y given C if every collider is in $an_{\mathcal{G}}(C)$ and no noncollider is in C.

Example

Is the walk $X_1 \to X_2 \leftarrow X_3 \to X_6$ d-connecting from X_1 to X_6 given $\{X_2, X_3\}$? Find a d-connecting walk between X_3 and X_4 given $\{X_2, X_6\}$.



Global Markov property

Definition (d-separation)

Let $A, B, C \subseteq V$ be disjoint sets. We say that A and B are *d-separated* if there is no d-connecting path given C between any $X \in A$ and any $Y \in B$, and we write $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$.

Note that d-separation is defined purely in terms of a graph, not the underlying SCM.

Theorem (Global Markov property, Proposition 6.31)

If P_X is induced by an SCM with graph $\mathcal{G}$, then for all disjoint $A, B, C \subseteq V$

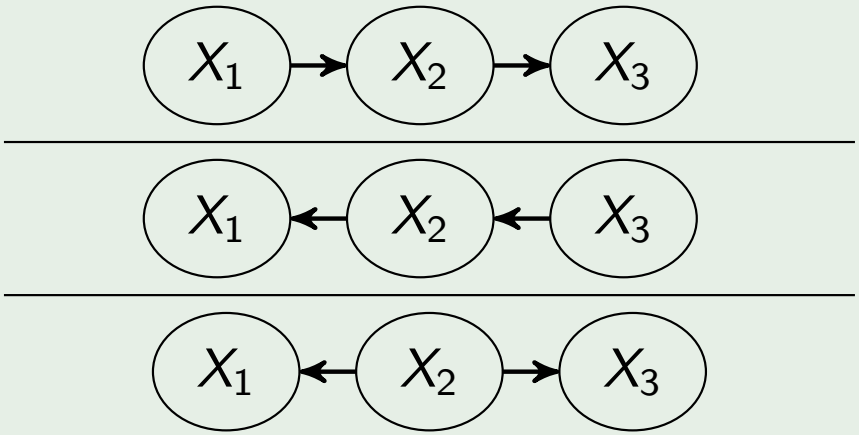
$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \Rightarrow A \perp\!\!\!\perp B \mid C.$$

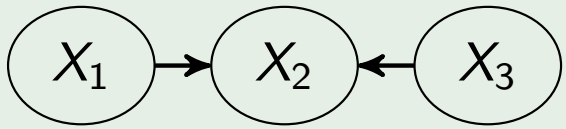
Global Markov property

Definition (d-separation)

Let $A, B, C \subseteq V$ be disjoint sets. We say that A and B are *d-separated* if there is no d-connecting path given C between any $X \in A$ and any $Y \in B$, and we write $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$.

Example



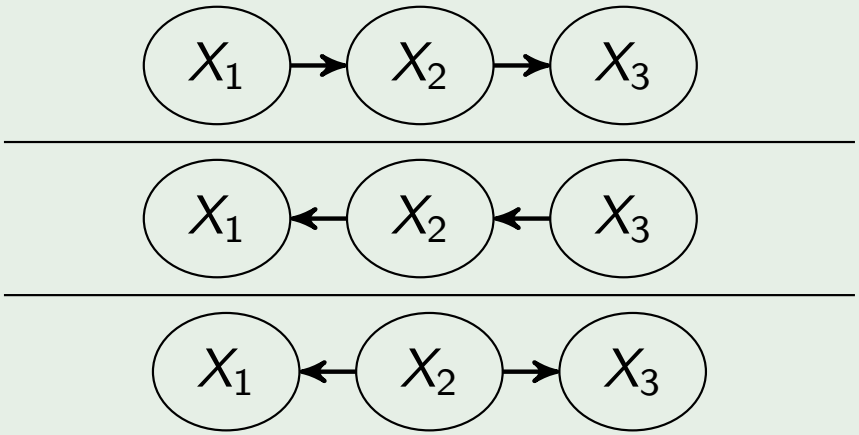


Global Markov property

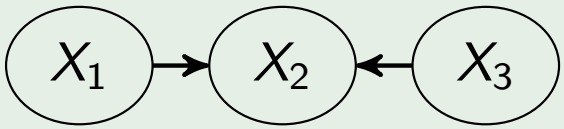
Definition (d-separation)

Let $A, B, C \subseteq V$ be disjoint sets. We say that A and B are *d-separated* if there is no d-connecting path given C between any $X \in A$ and any $Y \in B$, and we write $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$.

Example



$X_1 \perp\!\!\!\perp_{\mathcal{G}} X_3 \mid X_2$

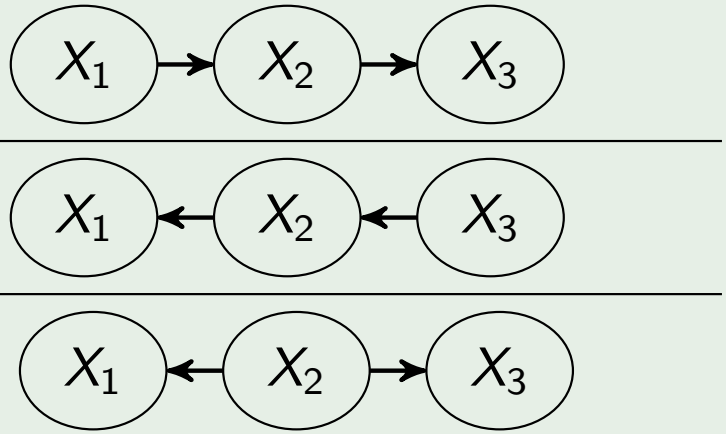


Global Markov property

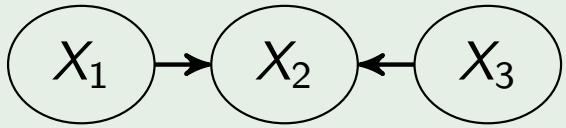
Definition (d-separation)

Let $A, B, C \subseteq V$ be disjoint sets. We say that A and B are *d-separated* if there is no d-connecting path given C between any $X \in A$ and any $Y \in B$, and we write $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$.

Example



$X_1 \perp\!\!\!\perp_{\mathcal{G}} X_3 \mid X_2$



$X_1 \perp\!\!\!\perp_{\mathcal{G}} X_3 \mid \emptyset$, that is, $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_3$

Existence of total causal effect (continued)

Proposition (Existence of total causal effect, Proposition 6.14)

- *If there is no directed path from X to Y, then there is no total causal effect from X to Y.*
- *There may be a directed path, but no total causal effect.*

We can prove the first statement easily using d-separation.

Existence of total causal effect (continued)

Proposition (Existence of total causal effect, Proposition 6.14)

- *If there is no directed path from X to Y, then there is no total causal effect from X to Y.*
- *There may be a directed path, but no total causal effect.*

We can prove the first statement easily using d-separation. We intervene on X and set it equal to some noise distribution, so there are no edges into X in the graph representing the interventional distribution. Any d-connecting path between X and Y must start with $X \to \ldots$

Existence of total causal effect (continued)

Proposition (Existence of total causal effect, Proposition 6.14)

- *If there is no directed path from X to Y, then there is no total causal effect from X to Y.*
- *There may be a directed path, but no total causal effect.*

We can prove the first statement easily using d-separation. We intervene on X and set it equal to some noise distribution, so there are no edges into X in the graph representing the interventional distribution. Any d-connecting path between X and Y must start with $X \to \dots$
Consider the empty conditioning set, $C = \emptyset$. A d-connecting path cannot have any colliders, so it must be directed. If there is no directed path from X to Y, then we have d-separation of X and Y given $\emptyset$ and therefore X and Y are independent by the global Markov property (regardless of $\tilde{N}_X$).

Causal questions

The SCMs are the fundamental building blocks of this course. Using these, we will ask different types of questions.

- Identification: With certain prior knowledge, can we identify a certain interventional distribution?
  - The truncated factorization shows that with knowledge of the graph and the observational distribution, we can identify the interventional distribution corresponding to a given intervention.
  - What if we only have partial observation (i.e., only a marginal distribution)?
  - What if we only have partial structural knowledge?
- Causal structure learning: Can we learn the graph from the observational and/or interventional distribution(s)?
  - How can we represent and understand partially observed causal systems?