# Causality
## Causal BNs and Structural Causal Models

Maciej Liśkiewicz

University of Lübeck

December, 2022

# Causal Models

Agenda

- *Causal* Bayesian Networks
- Modelling of controlled change in the model / controlled experiments
- Intervention
- *do*-operator
- Intervention joint distribution
- Causal effects
- Structural Causal Models / Functional Causal Models

# Causal Bayesian Networks

- So far, we have considered a BN as a carrier of conditional independence relationships
- However, such an "associational knowledge representation" is not sufficient when we want to make decisions under uncertainty
- Particularly, the interpretation of DAGs as carriers of CI assumptions does not necessarily imply causation

# Causal Bayesian Networks

- So far, we have considered a BN as a carrier of conditional independence relationships
- However, such an “associational knowledge representation” is not sufficient when we want to make decisions under uncertainty
- Particularly, the interpretation of DAGs as carriers of CI assumptions does not necessarily imply causation
- For example, in our “recommendation letter” BN, the (in)dependence between $I$ – student’s intelligence and $S$ – the student’s SAT score, can be encoded as

$$I \to S \quad \text{and} \quad I \leftarrow S$$

# Causal Bayesian Networks

- So far, we have considered a BN as a carrier of conditional independence relationships
- However, such an "associational knowledge representation" is not sufficient when we want to make decisions under uncertainty
- Particularly, the interpretation of DAGs as carriers of CI assumptions does not necessarily imply causation
- For example, in our "recommendation letter" BN, the (in)dependence between $I$ – student's intelligence and $S$ – the student's SAT score, can be encoded as

$$I \rightarrow S \quad \text{and} \quad I \leftarrow S$$

- Our goal: causal interpretation of DAGs
  - the ubiquity of graphical models in statistical and AI applications stems primarily from their causal interpretation

# Causal Bayesian Networks

- The advantages of building models around causal rather than associational information
  1. The judgments required in the construction of a DAG are more *meaningful* and therefore more *reliable*
  2. The ability to represent and respond to external or spontaneous *changes*
- Example: "slippery pavement" (Pearl (2009, Ch. 1))
- All these imply that causal models (assuming they are valid) are much more informative than probability models

# Causal Networks and Interventions

## Primacy of causal over associational knowledge...

**Example:** "Slippery Pavement" (Pearl (2009, ch. 1))

- $X_1$: the season of the year (spring, summer, fall, winter)
- $X_2$: whether rain falls
- $X_3$: whether the sprinkler is on
- $X_4$: whether the pavement would get wet
- $X_5$: whether the pavement would be slippery

$$
\begin{tikzcd}
& X_1 \ \text{season} \arrow[dl] \arrow[dr] & \\
\text{sprinkler} \ X_3 \arrow[dr] & & X_2 \ \text{rain} \arrow[dl] \\
& X_4 \ \text{wet} \arrow[d] & \\
& X_5 \ \text{slippery} &
\end{tikzcd}
$$

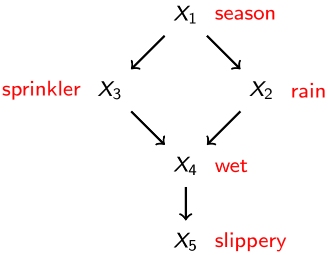
- ... meaningful judgments in the construction of a DAG

# Causal Networks and Interventions

## Primacy of causal over associational knowledge...

**Example:** "Slippery Pavement" (Pearl (2009, ch. 1))

- $X_1$: the season of the year (spring, summer, fall, winter)
- $X_2$: whether rain falls
- $X_3$: whether the sprinkler is on
- $X_4$: whether the pavement would get wet
- $X_5$: whether the pavement would be slippery

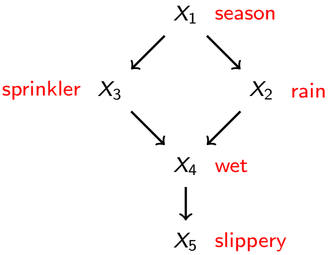



- ... meaningful judgments in the construction of a DAG
- **Exercise:** attempt to construct a DAG representation for the associations in DAG above along the ordering $(X_5, X_1, X_3, X_2, X_4)$

# Causal Networks and Interventions

Primacy of causal over associational knowledge...

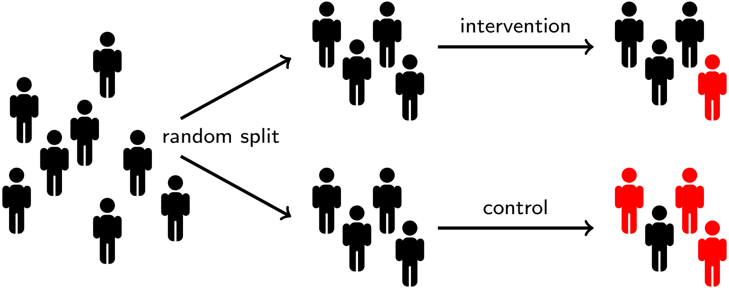
**Example** "Slippery Pavement" (Pearl (2009, ch. 1))




- ... the ability to represent and respond to external or spontaneous changes
- Each parent-child relationship represents a *stable* and *autonomous* physical mechanism
- It is conceivable to change one parent-child relationship without changing the others
- This property allows the following
  - A joint distribution tells us how probabilities would change with subsequent observations (as in case of BNs)
  - A causal model also tells us how these probabilities would change as a result of external interventions

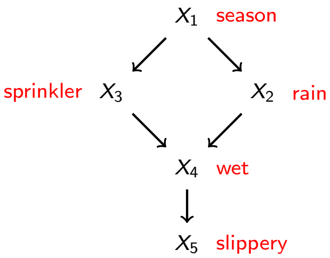
# Causal Networks and Interventions

- Direct experimentation



# Causal Networks and Interventions

- A causal model tells how the probabilities would change as a result of external interventions




- $P(x_1, x_2, x_3, x_4, x_5) \;=\; P(x_1)\; P(x_2 \mid x_1)\; P(x_3 \mid x_1)\; P(x_4 \mid x_2, x_3)\; P(x_5 \mid x_4)$
- In a causal model there is a deep connection between modularity and interventions
- Instead of specifying a new probability function $P$ for each of the many possible interventions, we specify merely the immediate change implied by the intervention
- As a result we get that the effect of an intervention can be predicted by modifying the corresponding factors in the decomposition $P(x_1, x_2, x_3, x_4, \ldots)$

# Causal Networks and Interventions

- A causal model tells how the probabilities would change as a result of external interventions


- Our "pre-intervention" probability function

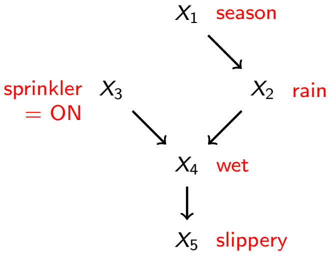
$$P(x_1, x_2, x_3, x_4, x_5) \;=\; P(x_1)\; P(x_2 \mid x_1)\; P(x_3 \mid x_1)\; P(x_4 \mid x_2, x_3)\; P(x_5 \mid x_4)$$

- Example: to represent the (external) action "turn the sprinkler On" we
  - remove $X_1 \to X_3$ from the model and
  - assign $X_3$ to the value "ON"
- We get the post-intervention probability function

$$P_{X_3=\mathrm{ON}}(x_1, x_2, x_4, x_5) \;=\; P(x_1)\; P(x_2 \mid x_1)\; P(x_4 \mid x_2, X_3 = \mathrm{ON})\; P(x_5 \mid x_4)$$

# Causal Networks and Interventions

- A causal model tells how the probabilities would change as a result of external interventions




- The post-intervention probability function

$$P_{X_3=\mathrm{ON}}(x_1, x_2, x_4, x_5) \;=\; P(x_1)\; P(x_2 \mid x_1)\; P(x_4 \mid x_2, X_3 = \mathrm{ON})\; P(x_5 \mid x_4)$$

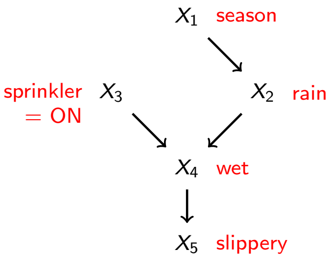
- The deletion of $P(x_3 \mid x_1)$ from the factorization means, whatever relationship existed between $X_1$ and $X_3$ prior to the action, that relationship is no longer valid while we perform the action

- We denote the action using the *do*-operator introduced by Pearl:

$$do(X_3 = \mathrm{ON})$$

# Causal Networks and Interventions

- A causal model tells how the probabilities would change as a result of external interventions




- We will use both notations (meaning the same)

$$P_{X_3=\mathrm{ON}}(x_1, x_2, x_4, x_5) \quad \text{and} \quad P(x_1, x_2, x_4, x_5 \mid do(X_3 = \mathrm{ON}))$$

  but we will prefer the last one

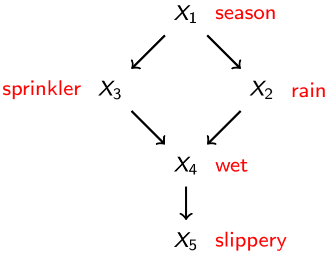
- According to the previously accepted notation this leads also to the following

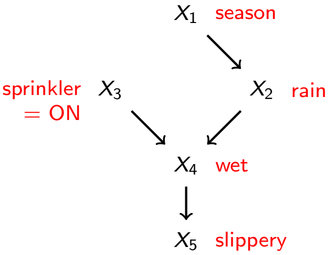
$$P_{x_3}(x_1, x_2, x_4, x_5) \quad \text{and} \quad P(x_1, x_2, x_4, x_5 \mid do(x_3))$$

  where by $x_3$ we mean a specific value of $X_3$ (in our case $x_3 = \mathrm{ON}$)

# Causal Networks and Interventions

- A causal model tells how the probabilities would change as a result of external interventions




- Note the difference between $P(x_1, x_2, x_4, x_5 \mid X_3 = \mathrm{ON})$ and $P(x_1, x_2, x_4, x_5 \mid do(X_3 = \mathrm{ON}))$
- $P(x_1, x_2, x_4, x_5 \mid X_3 = \mathrm{ON})$ is obtained by conditioning in the (left) model
- As a consequence: after observing that $X_3 = \mathrm{ON}$ (evidence), we wish to infer
  - that the season is dry
  - that it probably did not rain
  - etc.
- After the action "turning the sprinkler On" (righ) we infer quite different knowledge

# Causal Networks and Interventions

- Let $\mathbf{v}$ be a sequence of values of variables $V_1, \ldots, V_n$
- Let $\mathbf{X} \subseteq \mathbf{V} = \{V_1, \ldots, V_n\}$ and let $\mathbf{x}$ be a sequence of values of variables $\mathbf{X}$
- We say that $\mathbf{v} = (v_1, \ldots v_n)$ is consistent with $\mathbf{x} = (v'_{i_1}, \ldots, v'_{i_k})$ for all $i_j$ if $v_{i_j} = v'_{i_j}$

# Causal Networks and Interventions

- Let $\mathbf{v}$ be a sequence of values of variables $V_1, \ldots, V_n$
- Let $\mathbf{X} \subseteq \mathbf{V} = \{V_1, \ldots, V_n\}$ and let $\mathbf{x}$ be a sequence of values of variables $\mathbf{X}$
- We say that $\mathbf{v} = (v_1, \ldots v_n)$ is consistent with $\mathbf{x} = (v'_{i_1}, \ldots, v'_{i_k})$ for all $i_j$ if $v_{i_j} = v'_{i_j}$

## Interventional Distribution, Causal Effect

- Assume $G = (\mathbf{V} = \{V_1, \ldots, V_n\}, \mathbf{E})$ is a causal Bayesian network
- The interventional distribution $P(\mathbf{v} \mid do(\mathbf{x}))$ resulting from any intervention $do(\mathbf{X} = \mathbf{x})$ is defined as a truncated factorization:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{i:\ V_i \not\in \mathbf{X}\}} P(v_i \mid pa_i) \quad \text{for all } \mathbf{v} \text{ consistent with } \mathbf{x}$$

# Causal Networks and Interventions

- Let $\mathbf{v}$ be a sequence of values of variables $V_1, \ldots, V_n$
- Let $\mathbf{X} \subseteq \mathbf{V} = \{V_1, \ldots, V_n\}$ and let $\mathbf{x}$ be a sequence of values of variables $\mathbf{X}$
- We say that $\mathbf{v} = (v_1, \ldots v_n)$ is consistent with $\mathbf{x} = (v'_{i_1}, \ldots, v'_{i_k})$ for all $i_j$ if $v_{i_j} = v'_{i_j}$

## Interventional Distribution, Causal Effect

- Assume $G = (\mathbf{V} = \{V_1, \ldots, V_n\}, \mathbf{E})$ is a causal Bayesian network
- The interventional distribution $P(\mathbf{v} \mid do(\mathbf{x}))$ resulting from any intervention $do(\mathbf{X} = \mathbf{x})$ is defined as a truncated factorization:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{i:\ V_i \notin \mathbf{X}\}} P(v_i \mid pa_i) \quad \text{for all } \mathbf{v} \text{ consistent with } \mathbf{x}$$

- The causal effect of $\mathbf{X}$ on outcome variables $\mathbf{Y}$, denoted as

$$P(\mathbf{y} \mid do(\mathbf{x}))$$

  is defined as the probability distribution of variables $\mathbf{Y}$ after the intervention

# Causal Networks and Interventions

- Let $\mathbf{v}$ be a sequence of values of variables $V_1, \ldots, V_n$
- Let $\mathbf{X} \subseteq \mathbf{V} = \{V_1, \ldots, V_n\}$ and let $\mathbf{x}$ be a sequence of values of variables $\mathbf{X}$
- We say that $\mathbf{v} = (v_1, \ldots v_n)$ is consistent with $\mathbf{x} = (v'_{i_1}, \ldots, v'_{i_k})$ for all $i_j$ if $v_{i_j} = v'_{i_j}$

## Interventional Distribution, Causal Effect

- Assume $G = (\mathbf{V} = \{V_1, \ldots, V_n\}, \mathbf{E})$ is a causal Bayesian network
- The interventional distribution $P(\mathbf{v} \mid do(\mathbf{x}))$ resulting from any intervention $do(\mathbf{X} = \mathbf{x})$ is defined as a truncated factorization:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{i:\ V_i \notin \mathbf{X}\}} P(v_i \mid pa_i) \quad \text{for all } \mathbf{v} \text{ consistent with } \mathbf{x}$$

- The causal effect of $\mathbf{X}$ on outcome variables $\mathbf{Y}$, denoted as

$$P(\mathbf{y} \mid do(\mathbf{x}))$$

  is defined as the probability distribution of variables $\mathbf{Y}$ after the intervention

- Notice that if $\mathbf{v}$ is not consistent with $\mathbf{x}$, we let $P(\mathbf{v} \mid do(\mathbf{x})) = 0$

# Causal Networks and Interventions

**Example:** Observation vs. Intervention

- Assume $\mathbf{V} = \{R, W\}$
- $R$: it is raining ($R = 1$) or not ( $R = 0$);
- $W$: the street is wet ( $W = 1$ ) or not ($W = 0$)

# Causal Networks and Interventions

**Example:** Observation vs. Intervention

- Assume $\mathbf{V} = \{R, W\}$
- $R$: it is raining ($R = 1$) or not ( $R = 0$);
- $W$: the street is wet ( $W = 1$ ) or not ($W = 0$)

- Observation vs. Intervention $R$: it is raining ($R = 1$) or not ( $R = 0$); $W$: the street is wet ( $W = 1$ ) or not ($W = 0$)

$$R \longrightarrow W$$

# Causal Networks and Interventions

**Example:** Observation vs. Intervention

- Assume $\mathbf{V} = \{R, W\}$
- $R$: it is raining $(R = 1)$ or not $(R = 0)$;
- $W$: the street is wet $(W = 1)$ or not $(W = 0)$

$$R \longrightarrow W$$

- Suppose: $P(R = 1) = 0.01$, and $P(W = 1 \mid R = 1) = 1$, $P(W = 1 \mid R = 0) = 0.001$.
- Let's suppose we observe that the street is wet. Using Bayes Theorem, we get:

$$P(R = 1 \mid W = 1) = 0.91$$

# Causal Networks and Interventions

**Example:** Observation vs. Intervention

- Assume $\mathbf{V} = \{R, W\}$
- $R$: it is raining $(R = 1)$ or not $(R = 0)$;
- $W$: the street is wet $(W = 1)$ or not $(W = 0)$

$$R \qquad 1 = W$$

  - Suppose: $P(R = 1) = 0.01$, and $P(W = 1 \mid R = 1) = 1$, $P(W = 1 \mid R = 0) = 0.001$.
  - Let's suppose we observe that the street is wet. Using Bayes Theorem, we get:

$$P(R = 1 \mid W = 1) = 0.91$$

  - If we intervene to make the street wet, we get:

$$P(R = 1 \mid do(W = 1)) = 0.01$$

# Causal Networks and Interventions

**Example:** Observation vs. Intervention

- Assume $\mathbf{V} = \{R, W\}$
- $R$: it is raining $(R = 1)$ or not $(R = 0)$;
- $W$: the street is wet $(W = 1)$ or not $(W = 0)$

$$R \qquad 1 = W$$

- Suppose: $P(R = 1) = 0.01$, and $P(W = 1 \mid R = 1) = 1$, $P(W = 1 \mid R = 0) = 0.001$.
- Let's suppose we observe that the street is wet. Using Bayes Theorem, we get:

$$P(R = 1 \mid W = 1) = 0.91$$

- If we intervene to make the street wet, we get:

$$P(R = 1 \mid do(W = 1)) = 0.01$$

- We get, also that $P(R = 1 \mid do(W = 0)) = 0.01$
- Conclusion: $W$ has no causal effect on $R = 1$

# The Identifiability Problem

## Example: Smoking and the Genotype Theory (Pearl (2009))

- Does smoking ($S$) cause lung cancer ($C$)?
- Assume that the variables are binary, taking on true (1) or false (0) values
- No relevant factors are assumed
- Moreover, assume the following (hypothetical) data set from a study on the relations among cancer and cigarette smoking

| | Group Type | % of Population | % of Cancer cases |
|---|---|---|---|
| $S = 0$ | Nonsmokers | 50 | 9.75 |
| $S = 1$ | Smokers | 50 | 85.25 |

- Task: compute the causal effects $P(C = 1 \mid do(s))$ from data in this model

$$\underset{\text{Smoking}}{S} \longrightarrow \underset{\text{Cancer}}{C}$$

$$P(c \mid do(s)) = P(c \mid s)$$

- $P(C = 1 \mid do(S = 0)) \; = \; 0.0975$
- $P(C = 1 \mid do(S = 1)) \; = \; 0.8525$

## Identifiability of Causal Effects

The identification of the total causal effect (in general, non-parametric models), called the identification problem, is defined as follows:

- for a given DAG $G = (\mathbf{V}, \mathbf{E})$, a set $\mathbf{R} \subseteq \mathbf{V}$ representing observed variables, and two disjoint subsets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$,
- express the causal total effect $P(\mathbf{y} \mid do(\mathbf{x}))$ using only pre-intervention (i.e. do-operator free) probabilities involving variables in $\mathbf{R}$, or output that this is not possible.

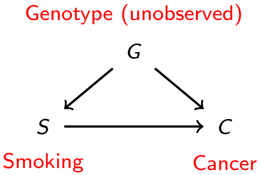
When such a formula exits, we say that the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ in $G$ is identifiable.

# The Identifiability Problem

## Example: Smoking and the Genotype Theory (Pearl (2009))

- Does smoking ($S$) cause lung cancer ($C$)?
- Assume, to forestall antismoking legislation, the tobacco industry has argued that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype that involves inborn craving for nicotine
- Thus, consider in our model the relevant factor: Genotype ($G$)
- Unfortunately, the feature is not measurable (called also unobserved)
- Can the causal effects $P(C = 1 \mid do(s))$ be estimated from data in this model?

| | Group Type | % of Population | % of Cancer cases |
|---|---|---|---|
| $S = 0$ | Nonsmokers | 50 | 9.75 |
| $S = 1$ | Smokers | 50 | 85.25 |




$P(c \mid do(s)) =$ not identifiable !

# The Identifiability Problem

## Example: Smoking and the Genotype Theory (Pearl (2009))

- Does smoking ($S$) cause lung cancer ($C$)?
- Consider now the amount of tar deposited in a person's lungs
- The relevant factors in our model are now: Genotype ($G$, unobserved), Tar in the lungs ($T$ binary, taking on true (1) or false (0) )
- How to compute the causal effects $P(C = 1 \mid do(s))$ from data:

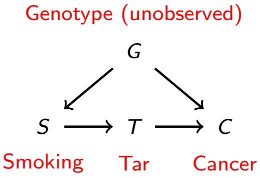
| | | Group Type | $P(s,t)$ % of Population | $P(C = 1 \mid s,t)$ % of Cancer cases |
|---|---|---|---|---|
| $S = 0$ | $T = 0$ | Nonsmokers, No tar | 47.5 | 10 |
| $S = 1$ | $T = 0$ | Smokers, No tar | 2.5 | 90 |
| $S = 0$ | $T = 1$ | Nonsmokers, Tar | 2.5 | 5 |
| $S = 1$ | $T = 1$ | Smokers, Tar | 47.5 | 85 |

# The Identifiability Problem

## Example: Smoking and the Genotype Theory (Pearl (2009))

- Task: compute the causal effects $P(C = 1 \mid do(s))$ from data:

| | | Group Type | $P(s, t)$ % of Population | $P(C = 1 \mid s, t)$ % of Cancer cases |
|---|---|---|---|---|
| $S = 0$ | $T = 0$ | Nonsmokers, No tar | 47.5 | 10 |
| $S = 1$ | $T = 0$ | Smokers, No tar | 2.5 | 90 |
| $S = 0$ | $T = 1$ | Nonsmokers, Tar | 2.5 | 5 |
| $S = 1$ | $T = 1$ | Smokers, Tar | 47.5 | 85 |



# The Identifiability Problem

## Example: Smoking and the Genotype Theory (Pearl (2009))

- Task: compute the causal effects $P(C = 1 \mid do(s))$ from data:

| | | Group Type | $P(s, t)$ % of Population | $P(C = 1 \mid s, t)$ % of Cancer cases |
|---|---|---|---:|---:|
| $S = 0$ | $T = 0$ | Nonsmokers, No tar | 47.5 | 10 |
| $S = 1$ | $T = 0$ | Smokers, No tar | 2.5 | 90 |
| $S = 0$ | $T = 1$ | Nonsmokers, Tar | 2.5 | 5 |
| $S = 1$ | $T = 1$ | Smokers, Tar | 47.5 | 85 |

Genotype (unobserved)

$$\begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S \arrow[r] & T \arrow[r] & C \end{tikzcd}$$

Smoking Tar Cancer

In our course we will show that the following formula can be used:

$$P(c \mid do(s)) = \sum_t P(t \mid s) \sum_{s'} P(c \mid s', t) P(s')$$

- $P(C = 1 \mid do(S = 0)) = .95(.10 \times .50 + .90 \times .50) + .05(.05 \times .50 + .85 \times .50)$
  $= .95 \times .50 + .05 \times .45 = .4975$
- $P(C = 1 \mid do(S = 1)) = .05(.10 \times .50 + .90 \times .50) + .95(.05 \times .50 + .85 \times .50)$
  $= .05 \times .50 + .95 \times .45 = .4525$

# Structural Causal Models

- Next we define more general causal models:
- Structural Causal Models (SCM) called also Functional Causal Models
- The first component of an SCM is a collection of assignments, e.g.
  - $x = f_X(u_X)$
  - $y = f_Y(x, u_Y)$

  that induces a DAG

$$\begin{tikzcd} U_X \arrow[d, dashed] & U_Y \arrow[d, dashed] \\ X \arrow[r] & Y \end{tikzcd}$$

- The second component is the probability function $P(U_X, U_Y)$

# Structural Causal Models

## Definition

### Structural Causal Model

- A structural causal model (SCM) $M = (S, P)$ consists of two sets of variables $\mathbf{V} = \{X_1, \dots, X_n\}$ and $\mathbf{U} = \{U_1, \dots, U_n\}$ and
  - a collection $S$ of $n$ (structural) assignments
    $$X_i = f_i(Pa_i, U_i)$$
    where $Pa_i \subseteq \mathbf{V}$ are called parents of $X_i$ and
  - a joint distribution $P$ over $\mathbf{U}$
- The variables in $\mathbf{U}$ are called exogenous variables, meaning that they are external to the model
- The variables in $\mathbf{V}$ are endogenous
- The graph $G$ associated with an SCM is obtained by creating one vertex for each $X_i$ and drawing directed edges from each parent in $Pa_i$ to $X_i$
- We assume the resulting $G$ is a DAG

# Structural Causal Models

## Basic Properties

- We sometimes call the exogenous variables in $\mathbf{U}$ the noise or error variables
- The elements of $Pa_i$ are called not only parents but also
  - direct causes of $X_i$, and
  - we call $X_i$ a direct effect of each of its direct causes

# Structural Causal Models

## Basic Properties

- We sometimes call the exogenous variables in $\mathbf{U}$ the noise or error variables
- The elements of $Pa_i$ are called not only parents but also
  - direct causes of $X_i$, and
  - we call $X_i$ a direct effect of each of its direct causes
- $P$ specifies the distribution of all exogenous variables
- In the case when the causal diagram is acyclic (as assumed in our definition), the corresponding model is called semi-Markovian

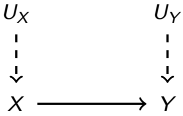
# Structural Causal Models

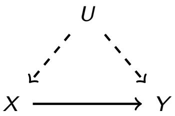
## Basic Properties

- We sometimes call the exogenous variables in $\mathbf{U}$ the noise or error variables
- The elements of $Pa_i$ are called not only parents but also
  - direct causes of $X_i$, and
  - we call $X_i$ a direct effect of each of its direct causes
- $P$ specifies the distribution of all exogenous variables
- In the case when the causal diagram is acyclic (as assumed in our definition), the corresponding model is called semi-Markovian
- In such models the values of the $X_i$ variables will be uniquely determined by those of the $\mathbf{U}$ variables
- Under such conditions, the joint distribution $P(x_1, \ldots, x_n)$ is determined uniquely by the distribution $P(\mathbf{u})$ of the error variables

# Structural Causal Models

- If, in addition to acyclicity, the error terms are jointly independent, the model is called Markovian
- Example of an Markovian and non-Markovian SCM




- Sometimes, in DAGs associated with Markovian SCMs we will not show explicitly the exogene variables $U_1, U_2, \ldots$
- By convention, this implies that they are assumed to be mutually independent.
- Additionally, we have the following properties
  - Every endogenous variable in a model is a descendant of at least one exogenous variable
  - Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable

# Structural Causal Models

## Basic Properties

### Theorem (Causal Markov Condition)

Assume $M = (S, P)$ is a Markovian SCM. Then

- $M$ defines a unique distribution over the variables $\mathbf{V} = \{X_1, \ldots, X_n\}$, such that $X_i = f_i(Pa_i, U_i)$ in distribution for $i = 1, \ldots, n$
- We refer to it as the associated with $M$ distribution and write $P_{\mathbf{V}}$
- The distribution $P_{\mathbf{V}}(x_1, \ldots, x_n)$ satisfies the parental Markov (local) condition relative the causal diagram $G$; that is, for each variable $X_i$:

$$(X_i \perp\!\!\!\perp \mathbf{V} \setminus (De(X_i) \cup Pa(X_i)) \mid Pa(X_i))_{P_{\mathbf{V}}}$$

# Structural Causal Models

## Basic Properties

### Theorem (Causal Markov Condition)

Assume $M = (S, P)$ is a Markovian SCM. Then

- $M$ defines a unique distribution over the variables $\mathbf{V} = \{X_1, \ldots, X_n\}$, such that $X_i = f_i(Pa_i, U_i)$ in distribution for $i = 1, \ldots, n$
- We refer to it as the associated with $M$ distribution and write $P_{\mathbf{V}}$
- The distribution $P_{\mathbf{V}}(x_1, \ldots, x_n)$ satisfies the parental Markov (local) condition relative the causal diagram $G$; that is, for each variable $X_i$:

$$(X_i \perp\!\!\!\perp \mathbf{V} \setminus (De(X_i) \cup Pa(X_i)) \mid Pa(X_i))_{P_{\mathbf{V}}}$$

$$\begin{tikzcd}
U_X \arrow[d, dashed] & U_Y \arrow[d, dashed] \\
X \arrow[r] & Y
\end{tikzcd}$$

**Exercise** Compute $P(x, y)$ in $M$, with $x = u_X,\ y = x \cdot u_Y + (1 - x)(1 - u_Y)$ where $U_X$ and $U_Y$ are two independent binary variables with $P(u_X = 1) = P(u_Y = 1) = 1/2$ (e.g., random coins)

## Definition (Intervention in SCMs, Causal Effect)

- Consider an SCM $M = (S, P)$ over $\mathbf{V} = \{X_1, \ldots, X_n\}$ and $\mathbf{U}$, with $S$ defined for $i = 1, \ldots, n$ as

$$X_i = f_i(Pa_i, U_i)$$

- To model the intervention

$$do(X_i = x_i), \quad \text{or in general} \quad do(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k})$$

  we replace one (or several) of the structural assignments $f$ for the variables to intervene, to obtain a new SCM $M' = (S', P)$, with

$$f_i'(Pa_i, U_i) := x_i$$

  and in general case

$$f_{i_1}'(Pa_{i_1}, U_{i_1}) := x_{i_1}, \ldots, f_{i_k}'(Pa_{i_k}, U_{i_k}) := x_{i_k}$$

- We define the causal effect of $\mathbf{X} \subseteq \mathbf{V}$ on outcome variables $\mathbf{Y} \subseteq \mathbf{V}$, denoted as

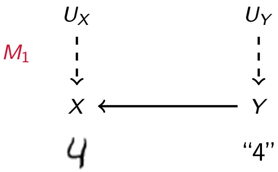
$$P_{\mathbf{V}}(\mathbf{y} \mid do(\mathbf{x}))$$

  as the probability distribution of variables $\mathbf{Y}$ after the intervention

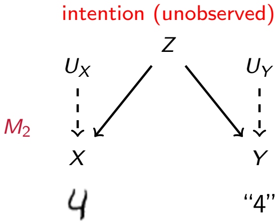
# Structural Causal Models

## Interventions and Causal Effects in SCM

**Example** (Pattern Recognition (Peters, Janzing, Schölkopf (2017), Sec. 1.4.1)
Structural causal models of handwritten digit data sets




$$X = f_X(Y, U_X)$$



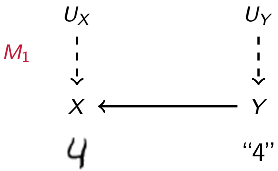

$$X = f_X(Z, U_X)$$
$$Y = f_Y(Z, U_Y)$$

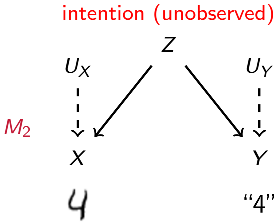
- Model $M_1$: a human is provided with class labels $y$ produces a corresponding handwritten digit image $x$; We model the process as a suitable function (or mechanism) $f_X$ of the class label $Y$ and some independent exogenous (noise) variable $U_X$
- We can then compute $P_{X,Y}$ from $P_{U_X,U_Y}$, and $f_X$
- There are two possible interventions in $M_1$, which lead to intervention distributions

# Structural Causal Models

## Interventions and Causal Effects in SCM

**Example** (Pattern Recognition (Peters, Janzing, Schölkopf (2017), Sec. 1.4.1)
Structural causal models of handwritten digit data sets




$$X = f_X(Y, U_X)$$

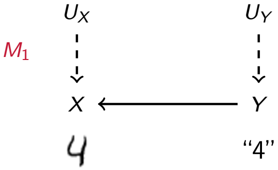
$$X = f_X(Z, U_X)$$
$$Y = f_Y(Z, U_Y)$$

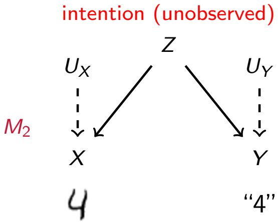
- Model $M_2$: the human decides which class to write ($Z$) and produces both images ($X$) and class labels ($Y$); Both the image $X$ and the recorded class label $Y$ are functions of the writer's intention ($Z$)
- We assume $Z$, $U_X$, and $U_Y$ are independent

# Structural Causal Models

## Interventions and Causal Effects in SCM

**Example** (Pattern Recognition (Peters, Janzing, Schölkopf (2017), Sec. 1.4.1)
Structural causal models of handwritten digit data sets




$$X = f_X(Y, U_X)$$

$$X = f_X(Z, U_X)$$
$$Y = f_Y(Z, U_Y)$$

- If the functions $f_X$ in $M_1$ and $f_X, f_Y$ in $M_2$ and noise terms are chosen suitably, we can ensure that $M_1$ and $M_2$ entail the same observational distributions $P_{X,Y}$
- However they are interventionally different

# Structural Causal Models vs Causal BN

## Linear SCMs (SEMs)

- In its general form, an SCM $M = (S, P)$ consists of a set $S$ of equations of the form
$$X_j = f_j(Pa_j, U_j) \quad j = 1, \ldots, n$$
- The equations are nonlinear, nonparametric generalization of the (recursive) linear structural equation models (SEMs)
$$X_j = \sum_{i<j} c_{ji} X_i + U_j, \quad j = 1, \ldots, n.$$
- Parameters $c_{ji}$ are called a path coefficients and they describe direct causal effects of $X_i$ on $X_j$
- Values $u_j$ represent error terms and there is assumed they have normal distribution

# Structural Causal Models vs Causal BN

## Linear SCMs (SEMs)

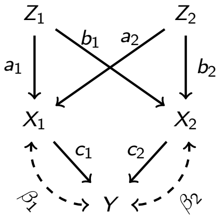
- In its general form, an SCM $M = (S, P)$ consists of a set $S$ of equations of the form
$$X_j = f_j(Pa_j, U_j) \quad j = 1, \ldots, n$$
- The equations are nonlinear, nonparametric generalization of the (recursive) linear structural equation models (SEMs)

$$X_j = \sum_{i<j} c_{ji} X_i + U_j, \quad j = 1, \ldots, n.$$

- Parameters $c_{ji}$ are called a path coefficients and they describe direct causal effects of $X_i$ on $X_j$
- Values $u_j$ represent error terms and there is assumed they have normal distribution
- **Example**

$$\begin{aligned}
Z_1 &= U_1 \\
Z_2 &= U_2 \\
X_1 &= a_1 Z_1 + a_2 Z_2 + U_3 \\
X_2 &= b_1 Z_1 + b_2 Z_2 + U_4 \\
Y &= c_1 X_1 + c_2 X_2 + U_5
\end{aligned}$$

$$\begin{aligned}
Cov(U_3, U_5) &= \beta_1 \neq 0 \\
Cov(U_4, U_5) &= \beta_2 \neq 0
\end{aligned}$$



# Structural Causal Models vs Causal BN

## Linear SCMs (SEMs)

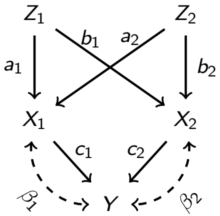
- In its general form, an SCM $M = (S, P)$ consists of a set $S$ of equations of the form
$$X_j = f_j(Pa_j, U_j) \quad j = 1, \ldots, n$$
- The equations are nonlinear, nonparametric generalization of the (recursive) linear structural equation models (SEMs)

$$X_j = \sum_{i<j} c_{ji} X_i + U_j, \quad j = 1, \ldots, n.$$

- Parameters $c_{ji}$ are called a path coefficients and they describe direct causal effects of $X_i$ on $X_j$
- Values $u_j$ represent error terms and there is assumed they have normal distribution
- **Example**

$$\begin{aligned}
Z_1 &= U_1\\
Z_2 &= U_2\\
X_1 &= a_1 Z_1 + a_2 Z_2 + U_3\\
X_2 &= b_1 Z_1 + b_2 Z_2 + U_4\\
Y &= c_1 X_1 + c_2 X_2 + U_5
\end{aligned}$$

$$\begin{aligned}
Cov(U_3, U_5) &= \beta_1 \neq 0\\
Cov(U_4, U_5) &= \beta_2 \neq 0
\end{aligned}$$




- SEM is a powerful multivariate analysis technique that is widely used by many applied researchers in the social and behavioral sciences

# Literature

- J. Pearl (2009), Ch.1
- J. Pearl, M. Glymour, and N.P. Jewell (2016), Ch. 1,2