Lecture 2: Causal Graphical Models

Tianpei Xie

Sep. 8th., 2022

Contents

1 Directed Graphical Models

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges, a *probabilistic graphical model (PGM)* defines a *joint probability distribution* $p(x_1, \ldots, x_m)$ that ***factorizes*** according to $\mathcal{G}$. Here each variable x_i corresponds to a node $v_i \in \mathcal{V}$ and the existence of an edge $(s,t) \in \mathcal{E}$ (or the ***absence*** of an edge (s,t)) defines a **statistical dependency** (or **conditional independency**) relation between x_s and x_t.

If the edge $(s,t) \in \mathcal{E}$ is **directed** (referred $s \to t$), i.e. there is a distinction between (s,t) and (t,s), the corresponding graphical models are referred as *directed graphical models*. Note that since a sequence of directed edges defines a logic flow, a circle in the graph would create undesireable contradiction. Therefore, we assume that $\mathcal{G}$ is a **directed acyclic graph (DAG)**, meaning that every edge is directed, and that the graph contains no directed cycles.

For any such DAG, we can define a **partial order** on the vertex set $\mathcal{V}$ by the notion of *ancestry*: we say that node s is an *ancestor* of u if there is a *directed path* $(s, t_1, t_2, \ldots, t_k, u)$. This is referred as ***topological ordering***. Given a DAG, for each vertex u and its parent set $\pi(u) = \{s \in \mathcal{V} : (s \to u) \in \mathcal{E}\}$, the conditional probability $p_s(x_s|x_{\pi(s)})$ is a **non-negative function** over $(x_s, x_{\pi(s)})$ and is **normalized** for all x_s, i.e. $\int p_s(x_s|x_{\pi(s)})dx_s = 1$.

The **directed graphical model** thus factorizes the joint distribution into a set of *local functions* $\{p_s(x_s|x_{\pi(s)}) : s \in \mathcal{V}\}$ according to the ancestor relations defined in $\mathcal{G}$

$$p(x_1, \ldots, x_m) = \prod_{s \in \mathcal{V}} p_s(x_s|x_{\pi(s)}). \tag{1}$$

This class of models are also referred as ***Bayesian networks*** [Koller and Friedman, 2009].

1.1 Conditional independence and d-separation

- **Definition** (***Pearl's d-separation***) [Koller and Friedman, 2009, Pearl, 2000, Peters et al., 2017]
  In a DAG $\mathcal{G}$, a ***path*** between nodes i_1 and i_m is ***blocked*** by a set S (with neither i_1 nor i_m in S) whenever *there is* a node i_k, such that **one** of the following two possibilities holds:

  1. $i_k \in S$ and

  $$\begin{aligned} & i_{k1} \to i_k \to i_{k+1} \\ \text{or} \quad & i_{k1} \leftarrow i_k \leftarrow i_{k+1} \\ \text{or} \quad & i_{k1} \leftarrow i_k \to i_{k+1} \end{aligned}$$

  2. neither i_k nor any of its descendants is in S and

  $$i_{k1} \to i_k \leftarrow i_{k+1}.$$

  (i_k is referred to as a ***collider***)

  Furthermore, in a DAG $\mathcal{G}$, we say that two **disjoint** subsets of vertices A and B are ***d-separated*** by a third (also **disjoint**) subset S if ***every path*** between nodes in A and B is ***blocked*** by S. We then write

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

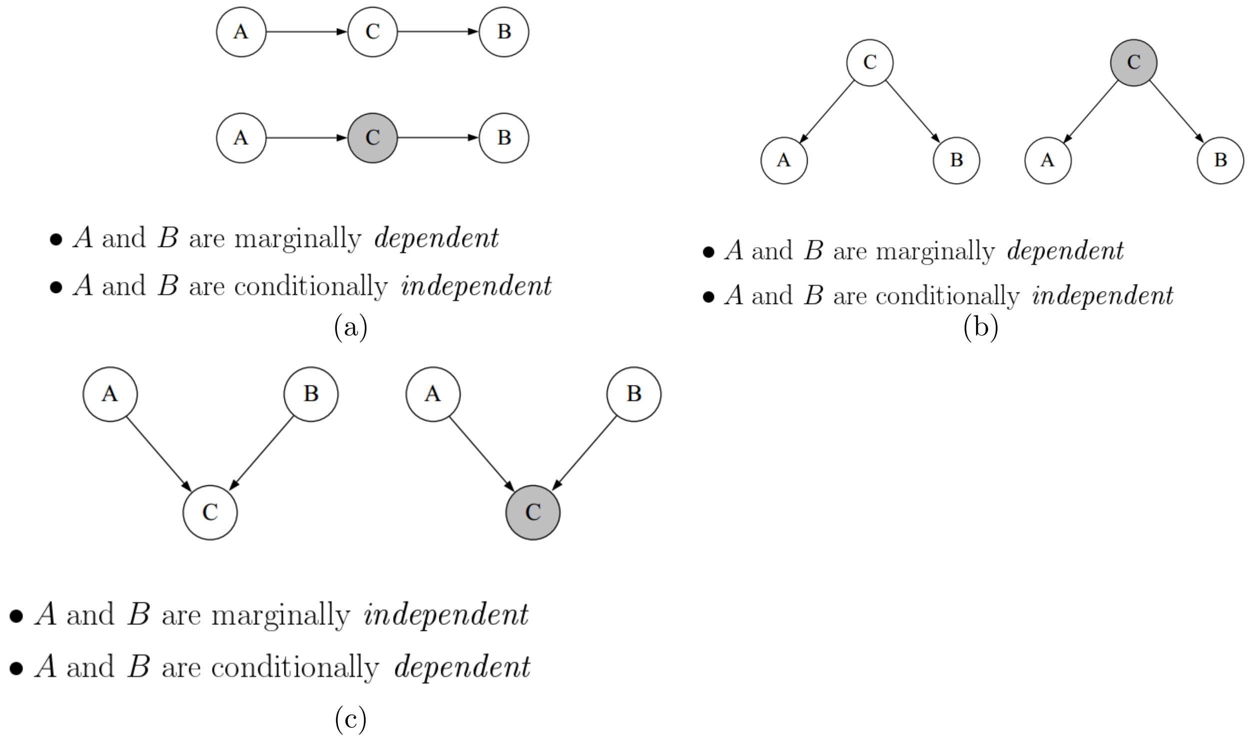

Figure 1: The conditional independency structure in directed graph. The shaded nodes are observed. In (c), the variable A and B are marginally *dependent* but conditionally *independent* due to collider C.

Note that the first possibility corresponds to Figure 1 (a) or (b) and the second possibility corresponds to Figure 1 (c).

- The above definition implies that the concept of path is replaced by the notion of ***active trail***, which not only consider directed path from $A \leftrightarrow B$ but also that of a v***-structure*** $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, when X_i or one of its descendant is in S. X_i is called a ***collider***.

  Definition [Koller and Friedman, 2009]
  Let $\mathcal{G}$ be a Bayesian network structure, and $X_1 \leftrightarrows \ldots \leftrightarrows X_n$ a *trail* in $\mathcal{G}$. Let Z be a subset of observed ***observed*** variable variables. The trail $X_1 \leftrightarrows \ldots \leftrightarrows X_n$ is ***active*** given Z if

  1. Whenever we have a *v-structure* $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its **descendants** are in Z;

  2. no other node along the trail is in Z.

  The existence of **v-structure dependency** indicates that sometimes dependent variables may not directly related but may related to a common variables. This is the effect of "**explaining away**". (like "an increase in activation of Earthquake leads to a decrease in activation of Burglar.") The trail from X_{i-1} to X_{i+1} becomes **active only when** the **collider** X_i is **observed**. That is why conditioning on descendant of a cause may introduce additional dependencies between cause and effect, which will break the exchangability.

- **Definition** We say that A and B are **d-separated** given S, denoted d-sep$_{\mathcal{G}}(A, B|S)$, if there is ***no active trail*** between any node $X \in A$ and $Y \in B$ given S.

The global Markov independence is

$$I(\mathcal{G}) = \left\{ (A \perp\!\!\!\perp B|S) : \text{ d-sep}_{\mathcal{G}}(A, B|S) \right\}$$

- The most important property in graphical models is the ***Markov properties*** over graph.

  Definition (***Markov property***) [Peters et al., 2017]
  Given a DAG $\mathcal{G}$ and a joint distribution $P_{\boldsymbol{X}}$, this distribution is said to satisfy

  1. the ***global Markov property*** with respect to the DAG $\mathcal{G}$ if

     $$A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, S \;\Rightarrow\; X_A \perp\!\!\!\perp X_B \,|X_S \tag{2}$$

     for all disjoint vertex sets A, B, S (the symbol $\perp\!\!\!\perp_{\mathcal{G}}$ denotes ***d-separation***)

  2. the ***local Markov property*** with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents,

     $$X_s \perp\!\!\!\perp X_{\mathcal{V}-\pi(s)-s} \,|\, X_{\pi(s)}$$

  3. the ***Markov factorization property*** with respect to the DAG $\mathcal{G}$ if

     $$p(x_1, \ldots, x_m) = \prod_{s \in \mathcal{V}} p_s(x_s | x_{\pi(s)}).$$

     For this last property, we have to assume that $P_{\boldsymbol{X}}$ has a density p; the factors in the product are referred to as *causal Markov kernels* describing the *conditional distributions* $P_{X_s|X_{\pi(s)}}$

- **Theorem 1.1** *(***Equivalence of Markov properties***)[Peters et al., 2017]*
  If $P_{\boldsymbol{X}}$ has a density p, then all Markov properties in Definition above are equivalent.

- The directed graphical model encodes a set of independency assertions.

  Definition [Koller and Friedman, 2009]
  Let $\mathcal{G}$ be any graph object associated with a set of independencies $I(\mathcal{G})$. We say that $\mathcal{G}$ is an ***I-map*** for a set of *independencies* I if $I(\mathcal{G}) \subseteq I$.

- Graphical model representation is not unique given a set of independencies I, i.e. there exists $\mathcal{G}_1 \neq \mathcal{G}_2$ so that $I(\mathcal{G}_1) = I(\mathcal{G}_2)$. These two graphs are ***I-equivalent***. We can define a *mininal representation* by proving that removing any of the edges in it will break the independency assertions in I.

  Definition A graph $\mathcal{G}$ is a ***minimal*** *I-map* for a set of independencies I if it is an I-map for I, and if the removal of even a single edge from $\mathcal{G}$ renders it not an I-map.

- **Definition** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. An ordering of the nodes $X_1, \ldots, X_n$ is a ***topological ordering*** relative to $\mathcal{G}$ if, whenever we have $X_i \to X_j \in \mathcal{E}$, then $i < j$.

1.2 Causal Graphical Model

- We introduce the concept of a causal structure.

Definition (***Causal Structure***) [Pearl, 2000]
A ***causal structure*** of a set of variables V is a **directed acyclic graph (DAG)** $\mathcal{G}$ in which each node corresponds to a distinct element of V, and each link represents a ***direct functional relationship*** among the corresponding variables.

A causal structure serves as a blueprint for forming a "causal model" a precise specification of how each variable is influenced by its parents in the DAG, as in the structural equation model.

- **Definition** (***Causal Model***) [Pearl, 2000]
  A **causal model** is a pair $M = (D, \Theta_D)$ consisting of a **causal structur**e D and a set of **parameters** Θ_D *compatible with* D. The parameters Θ_D assign a **function** $x_s = f_s(x_{\pi(s)}, u_s)$ to each $X_s \in V$ and a probability measure $P(u_s)$ to each u_s, where $X_{\pi(s)}$ are the parents of X_s in D and where each U_s is a random disturbance distributed according to $P(u_s)$, independently of all other u.

- **Definition** (***Latent Structure***) [Pearl, 2000]
  A **latent structure** is a pair $L = (D, O)$ where D is a causal structure over V and where $O \subseteq V$ is a set of observed variables.

- **Definition** (***Structure Preference***) One latent structure $L' = (D', O')$ is **preferred** to another $L = (D, O)$ (written $L' \succeq L$) if and only if D' can *mimic* D over O that is, if and only if for every Θ_D there exists a $\Theta_{D'}$ such that $P(O|D', \Theta_{D'}) = P(O|D, \Theta_D)$. Two latent structures are equivalent $L \equiv L'$ if and only if $L' \succeq L$ and $L \succeq L'$.

- **Definition** (***Consistency***) [Pearl, 2000]
  A latent structure $L = (D, O)$ is ***consistent*** with a distribution $\hat{P}$ over O if D can accommodate some model that generates $\hat{P}$ that is, if there exists a parameterization Θ_D such that $P(O|D, \Theta_D) = \hat{P}$.

- **Definition** (***Faithfulness*** and ***Causal Minimality***) [Peters et al., 2017]
  Consider a distribution $P_{\mathbf{X}}$ and a DAG $\mathcal{G}$.

  1. $P_{\mathbf{X}}$ is ***faithful*** to the DAG $\mathcal{G}$ if

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

     for all disjoint vertex sets A, B, C.

  2. A distribution satisfies ***causal minimality*** with respect to $\mathcal{G}$ if it is **Markovian** with respect to $\mathcal{G}$, but not to *any* proper *subgraph* of $\mathcal{G}$

- **Proposition 1.2** *(***Faithfulness implies causal minimality***) [Peters et al., 2017]*
  If $P_{\mathbf{X}}$ is faithful and Markovian with respect to $\mathcal{G}$, then causal minimality is satisfied.

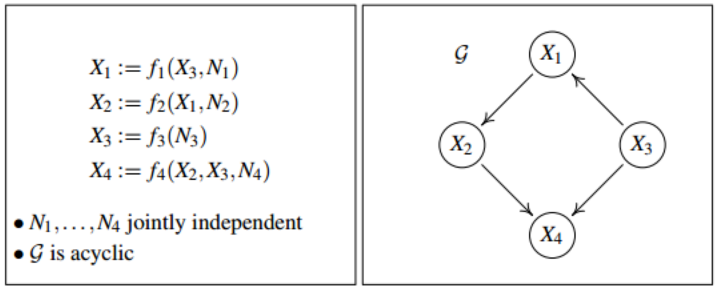

Figure 6.1: Example of an SCM (left) with corresponding graph (right). There is only one causal ordering π (that satisfies $3 \mapsto 1$, $1 \mapsto 2$, $2 \mapsto 3$, $4 \mapsto 4$).

Figure 2: The example of DAG created based on SCM [Peters et al., 2017]

2 Structural Causal Models

2.1 Definitions

- **Definition** (***Structural causal models (SCMs)***) [Peters et al., 2017]
  A **SCM** $\mathfrak{C} := (S, P_{\boldsymbol{N}})$ with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a collection S of d ***(structural) assignments***:

  $$X_s := f_s(X_{\pi(s)}, N_s), \quad s = 1, \dots, d \tag{3}$$

  where $X_{\pi(s)}$ are called **parents** of X_s; and a joint distribution $P_{\boldsymbol{N}} = \prod_{s=1}^{d} P_{N_s}$ over the noise variables, which we require to be ***jointly independent***.

  The $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a SCM is obtained by creating one vertex for each X_s and drawing **directed edges** from each parent in $X_{\pi(s)}$ to X_s, that is, from each variable X_k occurring on the right-hand side of equation (3) to X_s. $\mathcal{G}$ is a **directed acyclic graph (DAG)**.

  We sometimes call the elements of $X_{\pi(s)}$ not only parents but also ***direct causes*** of X_s, and we call X_s a ***direct effect*** of each of its direct causes. SCMs are also called (***nonlinear***) ***SEMs***.

  Structural assignments (3) should be thought of as a set of **assignments** or **functions** (rather than a set of mathematical equations) that tells us how certain variables determine others. This is the reason why we prefer to avoid the term ***structural equations***, which is commonly used in the literature.

- SCMs are the key for formalizing *causal reasoning* and *causal learning*. A SCM entails an **observational distribution**. But unlike usual probabilistic models, they additionally entail ***intervention distributions*** and ***counterfactuals***:

  Proposition 2.1 ***(Entailed distributions)*** *[Peters et al., 2017]*
  A SCM $\mathfrak{C}$ defines a ***unique*** *distribution over the variables $\boldsymbol{X} = (X_1, \dots, X_d)$ such that $X_s = f_s(X_{\pi(s)}, N_s)$, in distribution, for $s = 1, \dots, d$. We refer to it as the* ***entailed distribution*** *$P_{\boldsymbol{X}}^{\mathfrak{C}}$ and sometimes write $P_{\boldsymbol{X}}$.*

- In continous-time, we can rewrite the SCM in (3) as a set of *differential equations*. And the analysis on causality can be done at stationary state.

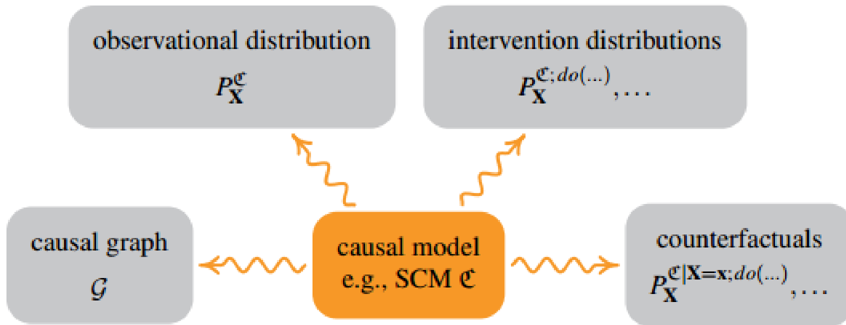

Figure 6.2: Causal models as SCMs do not only model an observational distribution P (Proposition 6.3) but also intervention distributions (Section 6.3) and counterfactuals (Section 6.4).

Figure 3: The structural cause models connect to observational, interventional and counterfactual analysis [Peters et al., 2017]

2.2 Cause-effect models

- A simple bivariate SCM is also called a **cause-effect model**.

  Definition (**Cause-Effect models**) [Peters et al., 2017]
  A **SCM** $\mathfrak{C}$ with graph $\mathcal{G} : C \to E$ consists of two assignments:

$$C := N_C, \tag{4}$$
$$E := f_E(C, N_E), \tag{5}$$

  where $N_E \perp\!\!\!\perp N_C$, that is, N_E is independent of N_C. Let $C \in \mathcal{C}$ and $E \in \mathcal{E}$ so that $f_E : \mathcal{C} \to \mathcal{E} \in \mathcal{E}^{\mathcal{C}}$.

- Note that from (5), E is deterministic given C and noise assignment n_E. Thus we can view n_E as choosing randomly from space of functions $\mathcal{E}^{\mathcal{C}}$. Based on this perspective, we can represent (5) in **canonical** form

$$E := N_E(C).$$

  This form implies that the choice of noise factor N_E determine the function $f_E \in \mathcal{E}^{\mathcal{C}}$.

- There are two types of *causal statements* entitled by SCM (4) and (5):

  1. The behavior of the system under **potential interventions**, i.e. $P_E^{do(C=c)} = P_{E|C=c}$.

     The ***interventional causal implications*** of the SCM are completely determined by the **marginal distributions** of each component of the vector-valued noise variable N_E even though the SCM includes a precise specification of P_{N_E}.

  2. The **counterfactual** statement. The counterfactual statements depend not only on the marginal distributions of the components of the noise variable N_E, but also on the statistical dependences between the outputs of functions $f_E \in \mathcal{E}^{\mathcal{C}}$ defined in (5).

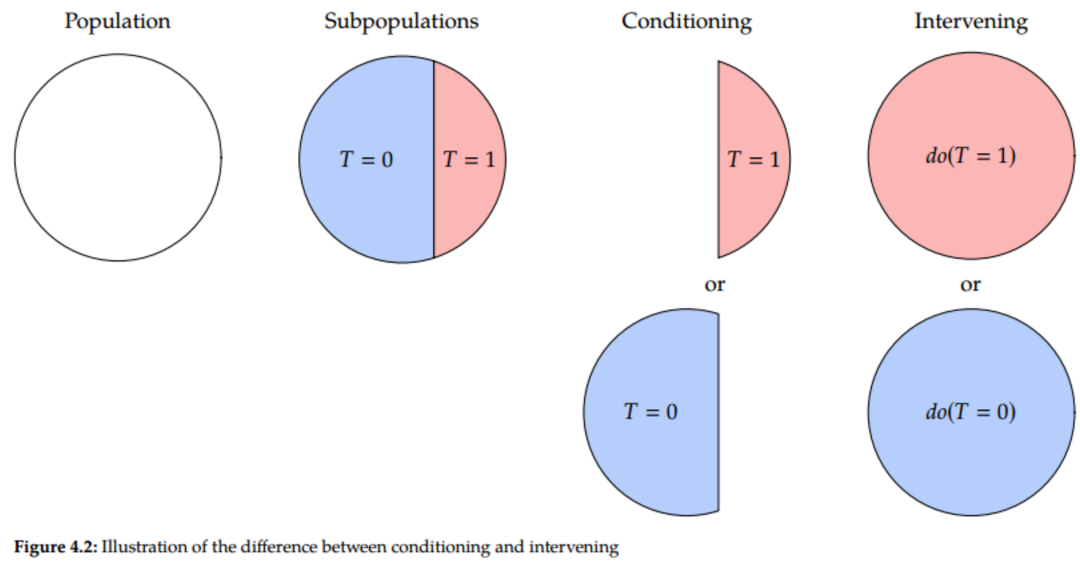

Figure 4: The condition and intervention are not the same [Neal, 2020]

2.3 Intervention

In causal inference, we are often interested in the systems behavior under an ***intervention***. The intervened system induces **another distribution**, which usually differs from the **observational distribution**. If any type of intervention can lead to an arbitrary **change of the system**, these two distributions become ***unrelated*** and instead of studying the two systems jointly we may consider them as *two separate systems*. This motivates the **idea** that after an intervention *only parts of the data-generating process change*.

The first thing that we will introduce is a mathematical operator for intervention, the ***do*-operator**. For example, when we replace the assignment (5) by $E := 4$. This is called a **(hard) intervention** and is *denoted* by $do(E := 4)$.

- **Definition** (***Intervention distribution***) [Peters et al., 2017]
  Consider an SCM $\mathfrak{C} := (S, P_{\boldsymbol{N}})$ and its entailed distribution $P_{\boldsymbol{X}}^{\mathfrak{C}}$. We ***replace*** one (or several) of the ***structural assignments*** to obtain a **new SCM** $\widehat{\mathfrak{C}}$. Assume that we replace the assignment for X_k by

  $$\widehat{X}_k := \widehat{f}_k(\widehat{X}_{\pi(k)}, \widehat{N}_k).$$

  We then call the entailed distribution of the new SCM an ***intervention distribution*** and say that the variables whose structural assignment we have replaced have been **intervened** on. We denote the new distribution by

  $$P_{\boldsymbol{X}}^{\widehat{\mathfrak{C}}} := P_{\boldsymbol{X}}^{\mathfrak{C}; do\left(\widehat{X}_k := \widehat{f}_k(\widehat{X}_{\pi(k)}, \widehat{N}_k)\right)}.$$

  The set of noise variables in $\widehat{\mathfrak{C}}$ now contains both some "new" $\widehat{N}$s and some "old" Ns, all of which are required to be **jointly independent**.

  When $\widehat{f}_k(\widehat{X}_{\pi(k)}, \widehat{N}_k)$ puts a point mass on a real value a, we simply write $P_{\boldsymbol{X}}^{\mathfrak{C}; do(X_k := a)}$ and call this an ***atomic intervention***. An intervention with $\widehat{X}_{\pi(k)} = X_{\pi(k)}$, that is, where direct causes *remain* direct causes, is called ***imperfect***. This is a special case of a ***stochastic intervention*** [Korb et al., 2004], in which the marginal distribution of the intervened variable has positive variance.

We require that the new SCM $\widehat{\mathfrak{C}}$ have an **acyclic graph** $\widehat{\mathcal{G}}_{X_k}$; the set of allowed interventions thus depends on the graph induced by $\mathfrak{C}$.

2.3.1 *do*-operators

- The ***do*-operator** is different from **conditioning**. Conditioning on $T = t$ just means that we are restricting our *focus* to the ***subset*** of the population to those who have treatment $T = t$. In contrast, an intervention would be to take ***the whole population*** and give everyone treatment $T = t$.

- The notation of *do*-operator is commonly used in graphical causal models, and it has equivalents in **potential outcomes** notation.

$$P(Y(t) = y) := P(Y = y \,|\, do(T = t)) := P(y \,|\, do(t)) := P_Y^{do(T=t)} \tag{6}$$

  The ***ATE (average treatment effect)*** can be written as

$$\mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}\left[Y \,|\, do(T = 1)\right] - \mathbb{E}\left[Y \,|\, do(T = 0)\right] \tag{7}$$

  As discussed above, the distribution $P_Y^{do(T=t)}$ is not conditional distribution $P_{Y|T=t}$ but full distribution of Y under intervention $T = t$. Denote the corresponding density of interventional distribution $p^{do(T=t)}(c)$.

- If an expression Q with *do*-operator can be converted to without *do* in it, this expression is ***identifiable***. We will refer to an *estimand* as a ***causal estimand*** when it contains a *do*-operator, and we refer to an estimand as a ***statistical estimand*** when it doesnt contain a *do*-operator.

- Whenever, $do(T = t)$ appears **after** the conditioning bar, it means that everything in that expression is in the ***post-intervention*** world where the intervention $do(T = t)$ occurs.

  For example $\mathbb{E}\left[Y|do(T = t), Z = z\right]$ refers to the expected outcome in the subpopulation where $Z = z$ after the whole subpopulation has taken treatment $T = t$. On the other hand, $\mathbb{E}\left[Y|Z = z\right]$ simply refers to the expected value in the (*pre-intervention*) population where individuals take whatever treatment they would normally take (T)).

- Instead of hard intervention, we can have $do(E = g_E(C) + \hat{N}_E)$, which keeps a functional dependence on C but changes the noise distribution. This is an example of a **soft intervention**.

- Intervention on effect variables E will **not** change the distribution of cause variables C. $P_C^{do(E=e)} = P_C^{\mathcal{C}}$ for all e. On the other hand, intervention on the "cause" variables C will change the distribution of "effect" variables E. For instance, in (4) and (5) let N_E and N_C be standard normal distributed $N(0, 1)$. Let $E = 4C + N_E$. Then $P_E^{\mathcal{C}} = N(0, 17) \neq P_E^{do(C=2)} = N(8, 1) = P_{E|C=2}^{\mathcal{C}}$.

  The **asymmetry** between cause and effect can also be formulated as an **independence statement**. Intervention on effect variables E will break the dependency between C and E so that $(C \perp\!\!\!\perp E)_{do(E=e)}$ under intervention.

2.3.2 *do*-calculus

- **Proposition 2.2** ***(Rules of do Calculus)****[Pearl, 2000]*
  Let $\mathcal{G}$ be the directed acyclic graph associated with a causal model as defined in (4), (5)*, and let $P(\cdot)$ stand for the probability distribution induced by that model. For any* ***disjoint*** *subsets of variables X, Y, Z, and W, we have the following rules.*

  1. *(**Insertion/deletion of observations**):*
  $$p(y|\hat{x}, z, w) = p(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_X} \tag{8}$$
  where $\hat{x} := do(X = x)$ and $\widehat{\mathcal{G}}_X$ is induced sub-graph under intervention $\hat{x}$.

  2. *(**Action/observation exchange**):*
  $$p(y|\hat{x}, \hat{z}, w) = p(y|\hat{x}, z, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_{X,Z}} \tag{9}$$
  where $\widehat{\mathcal{G}}_{X,Z}$ is induced sub-graph under intervention $\hat{x}, \hat{z}$.

  3. *(**Insertion/deletion of actions**):*
  $$p(y|\hat{x}, \hat{z}, w) = p(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_{X,Z(W)}} \tag{10}$$
  where $Z(W)$ is the set of Z-nodes that are ***not ancestors*** *of any W-node in $\widehat{\mathcal{G}}_X$.*

  The equation (8) is based on d-separation of graphical model **after the intervention** $do(X = x)$. The equation (9) provides a condition for an ***external intervention*** $do(Z = z)$ to have the same effect on Y as the ***passive observation*** $Z = z$. (10) provides conditions for *deleting* (or introducing) an external intervention $do(Z = z)$ introduces no new association that can affect the probability of $Y = y$. $Z(W)$ is an additional constraint set to prevent inducing association by conditioning on the descendants of *colliders*.

- **Theorem 2.3** *(**Do-calculus are complete**) [Peters et al., 2017]*
  The following statements hold.

  1. *The rules are* ***complete****; that is, all identifiable intervention distributions can be computed by an iterative application of these three rules [Huang and Valtorta, 2006, Shpitser and Pearl, 2006];*

  2. *In fact, there is an algorithm that is guaranteed [Huang and Valtorta, 2006, Shpitser and Pearl, 2006] to find all identifiable intervention distributions.*

  3. *There is a necessary and sufficient graphical criterion for identifiability of intervention distributions [Shpitser and Pearl, 2006], based on so-called hedges [Huang and Valtorta, 2006].*

2.4 Counterfactuals

Another possible modification of an SCM changes all of its noise distributions. Such a change can be induced by observations and allows us to answer **counterfactual questions** such as "What if i did this, what would the outcome be ?". The **counterfactual outcome** is the result of **intervention on alternative cause** in SCM ***given the observation of current cause and effect***.

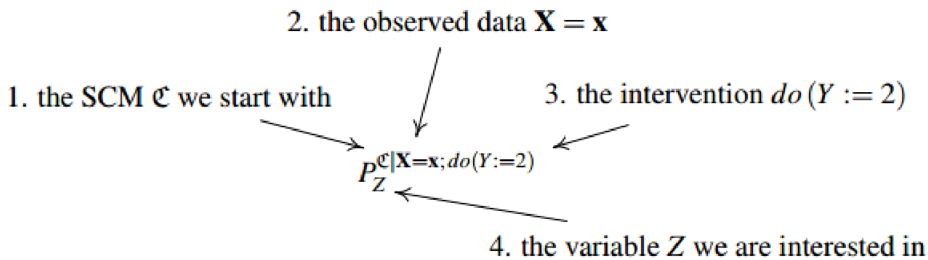

Figure 5: The diagram illustration of steps in counterfactual analysis. [Peters et al., 2017]

- **Definition** (***Counterfactuals***) [Peters et al., 2017]
  Consider an SCM $\mathfrak{C} := (S, P_{\boldsymbol{N}})$ over nodes $\boldsymbol{X}$. Given some observations $\boldsymbol{x}$, we define a ***counterfactual*** SCM by **replacing** the distribution of ***noise variables***:

$$\mathfrak{C}_{\boldsymbol{X}=\boldsymbol{x}} = \left(S, P_{\boldsymbol{N}}^{\mathfrak{C}|\boldsymbol{X}=\boldsymbol{x}}\right)$$

  where $P_{\boldsymbol{N}}^{\mathfrak{C}|\boldsymbol{X}=\boldsymbol{x}} = P_{\boldsymbol{N}|\boldsymbol{X}=\boldsymbol{x}}$. The new set of noise variables **need not be jointly independent** anymore. Counterfactual statements can now be seen as *do*-statements in the *new counterfactual SCM*.

  This definition can be generalized such that we observe not the full vector $\mathrm{X} = \mathrm{x}$ but only some of the variables.

- The steps for counterfactual analysis in Figure 5:
  1. Given SCM $\mathfrak{C}$, we can first **condition** on *observations* $\boldsymbol{X} = \boldsymbol{x}$ to update the distribution over the noise variables.
  2. Next, we calculate the effect of **intervention** on alternative treatment $do(Y := 2)$ for the SCM.
  3. Finally, we can compute the probablilty of outcome Z conditioned on both observations and do-operator, i.e. $P_Z^{\mathfrak{C}|\boldsymbol{X}=\boldsymbol{x},do(Y=2)}$ as the counterfactural outcome.

- Counterfactual statements depend strongly on the **structure of the SCM**. Two SCMs can induce the *same graph*, *observational distributions*, and *intervention distributions* but *entail* ***different counterfactual statements***. We will call those SCMs "*probabilistically and interventionally equivalent*" but not "*counterfactually equivalent*".

  In this sense, causal graphical models are *not* ***rich*** *enough* to predict **counterfactuals**.

  Definition (***Equivalence of causal models***) [Peters et al., 2017]
  Two models are called

  {probabilistically / interventionally / counterfactually} equivalent

  if they entail the same {obs. / obs. **and** int. / obs., int., **and** counter.} distributions.

2.5 Truncated Factorization

- **Proposition 2.4** *(***Truncated Factorization***) [Pearl, 2000, Peters et al., 2017, Neal, 2020] We assume that p and $\mathcal{G}$ satisfy the Markov assumption and modularity. Given, a set of* ***intervention*** *nodes S, if $\boldsymbol{x}$ is* ***consistent*** *with the intervention, then*

$$p(x_1, \ldots, x_m \mid do(S = s)) = \prod_{i \notin S} p_i(x_i | x_{\pi(i)}) \tag{11}$$

  Otherwise, $p(x_1, \ldots, x_m | do(S = s)) = 0$.

  That is for all factors related to $X_i \in S$, the values are set to be 1 due to intervention. In other words, the factors for (11) have been *truncated* compared to (1).

- An *alternative interpretation* is that for any SCM $\widehat{\mathfrak{C}}$ obtained from $\mathfrak{C}$ by intervening on some X_i, we have the following *invariance* statement:

$$p^{\widehat{\mathfrak{C}}}(x_j \mid x_{\pi(j)}) = p^{\mathfrak{C}}(x_j \mid x_{\pi(j)}), \quad \forall\, j \neq i. \tag{12}$$

  The equation in (12) shows that *causal relationships* are ***autonomous*** under interventions: if we intervene on a variables, the other mechanisms remain ***invariant***.

- Truncated factorization is also called ***G-computation formula*** [Imbens and Rubin, 2015] and **manipulation theorem**.

3 The Principle of Independent Mechanisms

Given two variables A, T and their joint distribution $p(a, t)$, how to determine the causal structure ($A \to T$ or $T \to A$) ? A first idea is to consider the **effect of interventions**. If we can change the value of A, how would the value of T change ? Here we assume that *the physical mechanism* $p(t|a)$ responsible for producing T given A. If $A \to T$ is a causal relationship, this would hold true independent of the distribution from A, $p(a)$.

Specifically, if $A \to T$ is the correct causal structure, then

1. it is in principle possible to perform a ***<u>localized intervention</u>*** on A, in other words, to change $p(a)$ without changing $p(t|a)$, and
2. $p(a)$ and $p(t|a)$ are ***<u>autonomous</u>***, ***<u>modular</u>***, or ***invariant*** mechanisms or objects in the world.

- In the **causal factorization** $p(a, t) = p(t|a)p(a)$, we would expect that the *conditional density* $p(t|a)$ (viewed as a ***function*** of t and a) provides no information about the ***marginal*** *density function* $p(a)$. This holds true if $p(t|a)$ is a model of a ***physical mechanism*** that does not care about what distribution $p(a)$ we feed into it. This is called the **independence of cause and mechanism**.

- In previous example, we can write $A \to T$ into a SCM

$$\begin{aligned} A &= N_A \\ T &= f_T(A, N_T) \end{aligned}$$

  where N_T and N_A are **statistically independent noises** $N_T \perp\!\!\!\perp N_A$.

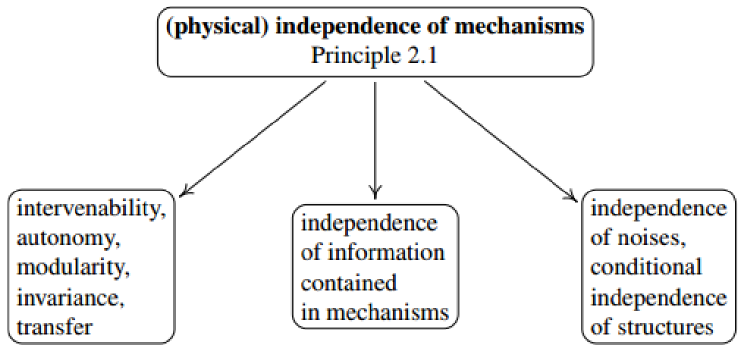

Figure 2.2: The principle of independent mechanisms and its implications for causal inference (Principle 2.1).

Figure 6: The independent mechanism priniciple [Peters et al., 2017]

- **Principle 3.1** ***(Independent mechanisms)*** *[Peters et al., 2017]*
  The causal generative process of a systems variables is composed of ***autonomous modules*** *that do not inform or influence each other.*

  In the probabilistic case, this means that the ***conditional distribution*** *of each variable given its* ***causes*** *(i.e., its mechanism) does* ***not inform or influence*** *the other conditional distributions. In case we have only two variables, this reduces to an* ***independence*** *between the* ***cause*** *distribution and the* ***mechanism*** *producing the effect distribution.*

- **Assumption 3.2** ***(Ignorability / Exchangeability)*** *[Neal, 2020]*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \tag{13}$$

  Note that $(Y(1), Y(0))$ describes the mechansim and $P_{Y(t)} = P_Y^{do(T=t)} = P_{Y|T=t}$ when $T \to Y$ is a causal structure. This assumption is the same as the independent mechanisms principle.

- The independent mechanism assumption implies that we can **change one mechanism without affecting the others**, – or, in causal terminology, we can **intervene** on one mechanism without affecting the others. An assumption such as this one is often implicit to **justify** the possibility of interventions in the first place, but one can also view it as a more general basis for causal reasoning and causal learning.

- The existence of an **invariant** mechanism under local intervention can be used in domain adapation and transfer learning [Peters et al., 2017].

- It is important to distinguish between two levels of information: an **effect** contains information about its cause, but the **mechanism** that ***generates*** the effect from its cause contains *no* information about the mechanism generating the cause. $(P_{Y(1),Y(0)} = P_Y^{do(T=0,1)} \neq P_Y)$

4 Controlling confounding bias

- ***Covariates*** or ***confounding factors*** are variables other than the cause and effect variables of interest but have impact on them.

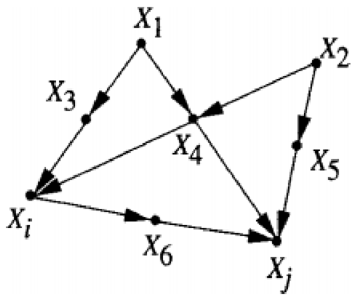

Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 7: The back-door adjustment [Pearl, 2000]

Definition (***Confounding***) [Peters et al., 2017]
Consider an SCM $\mathfrak{C}$ over nodes $\mathcal{V}$ with a directed path from $X \to Y$, $X, Y \in \mathcal{V}$. The causal effect from X to Y is called ***confounded*** if

$$p^{\mathfrak{C}:do(X=x)}(y) \neq p^{\mathfrak{C}}(y|x). \tag{14}$$

Otherwise, the causal effect is called ***unconfounded***.

- In order to account for the influence of confounder, we should **partition** the population into groups that are ***homogeneous*** relative to *confounder* Z, assessing the effect of X on Y in each homogeneous group, and then averaging the results. This process is called ***covariate adjustment***. This is the idea behind the ***Adjustment Formula*** [Imbens and Rubin, 2015].
- ***Confounder*** should be distinguished from the ***collider***: confounders ***need*** to be **controlled for** when estimating causal associations, while collider should be **avoided** during the conditioning.

This section discuss the process of **choosing** adjustment set using causal structure.

4.1 The Back-door Adjustment

Assume we are given a ***causal diagram*** $\mathcal{G}$, together with *nonexperimental* data on a subset V of observed variables in $\mathcal{G}$, and suppose we wish to estimate what effect the interventions $do(X = x)$ would have on a set of response variables Y, where X and Y are two subsets of V. In other words, we seek to estimate $P(y \mid do(x))$ from a sample estimate of $P(v)$, given the assumptions encoded in $\mathcal{G}$.

The ***back-door adjustment*** or *back-door criterion* [Pearl, 2000] is a simple graphical test that can be applied directly to the causal diagram in order to test if a set $Z \subseteq V$ of variables is sufficient for identifying $P(y \mid do(x))$.

- **Definition** (**Back-Door**) [Pearl, 2000]
  A set of variables Z satisfies the ***back-door criterion*** relative to an ***ordered*** pair of variables $(X_i \to X_j)$ in a DAG $\mathcal{G}$ if:

  1. ***no*** node in Z is a ***descendant*** of X_i; and
  2. Z ***blocks every path*** between X_i and X_j that contains an arrow ***into*** X_i.

  Similarly, if X and Y are two **disjoint** subsets of nodes in $\mathcal{G}$, then Z is said to satisfy ***the***

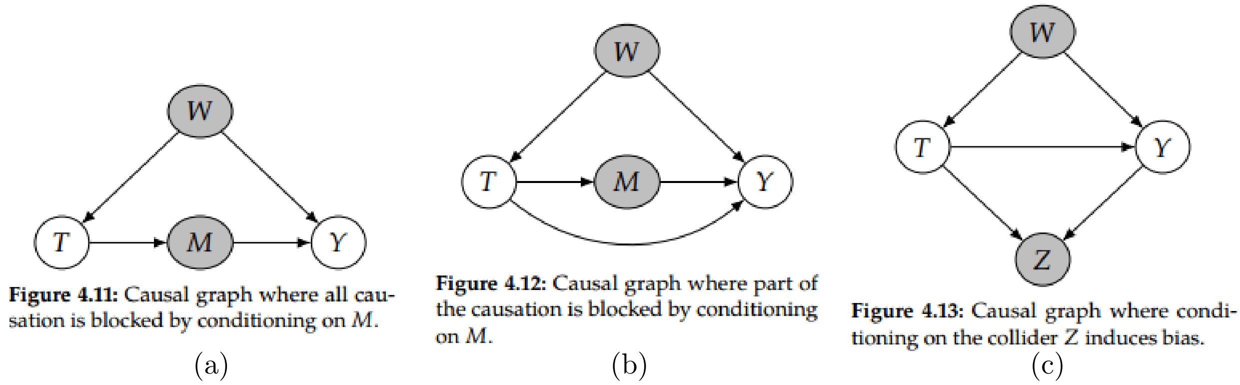

Figure 8: The collider bias induced by conditioning on descendant of treatment [Neal, 2020]

back-door criterion relative to (X, Y) if it satisfies the criterion relative to *any pair* (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

- The first condition makes sure ***no descendant of the treatment is included***. The second condition requires that *only paths with arrows* ***pointing at*** X_i *be blocked*; these paths can be viewed as **entering** X_i through the ***back door***.
- Satisfying the back-door criterion makes Z a ***sufficient adjustment set***. The main insight of the graphical approach to covariate adjustment is that the adjustment set must **block all *noncausal* paths without blocking** any ***causal*** paths between X and Y.
- X and Y are **not d-separated** given Z since the *front-door path* $X \rightarrow Y$ is not blocked.
- **Theorem 4.1** *(**Back-Door Adjustment**) [Pearl, 2000, Neal, 2020]*
  *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is **identifiable** and is given by the formula*

  $$P(y \,|\, do(x)) = \sum_{z} P(y \,|\, x, z) P(z) \tag{15}$$

  To see why this works we need to know that $P(z|do(x)) = P(z)$ since by back-door criterion, Z has no descendant of X. Also $P(y \,|\, do(x), z) = P(y \,|\, x, z)$ since Z blocks all paths from X to Y, so by modularity
- The summation in (15) represents the standard formula obtained under adjustment for Z; variables X for which the equality in (15) is valid were named "***conditionally ignorable*** *given* Z" as in [Imbens and Rubin, 2015].

  $$(Y(1), Y(0)) \perp\!\!\!\perp X \mid Z$$
- Using back-door criterion, we can choose set Z that blocks the non-collider path from Y to X.

4.2 Collider Bias and Why to Not Condition on Descendants of Treatment

There are two categories of things that could go wrong if we condition on descendants of treatment X:

1. We **block the flow of causation** from X to Y.

   If we condition on a node that is on a directed path from X to Y, then we block the flow of causation along that causal path. We will refer to a node on a directed path from X to Y as a ***mediator***, as it mediates the effect of X on Y. This way we have either completely independent variables conditioned on mediator Z (Figure 8 (a)) or we have biased estimation (Figure 8 (b)).

2. We **induce non-causal association** between X and Y if Z is a collider, i.e. $X \to Z \leftarrow Y$.

   If we condition on a descendant of X that isnt a mediator, it could ***unblock*** a path from X to Y that was blocked by a collider (Figure 8 (c)). Conditioning on Z, or any descendant of Z in a path like this, will induce ***collider bias***. That is, the causal effect estimate will be biased by the non-causal association that we induce when we condition on Z or any of its descendants.

4.3 Compare to Adjustment Formula

From above we can see that the Adjustment formula from Potential Outcome thoery is equivalent to the Back-Door Adjustment from SCMs.

$$\begin{aligned}
\mathbb{E}\left[Y_i(x)\right] = \mathbb{E}\left[Y|do(x)\right] &= \mathbb{E}_Z\left[\mathbb{E}\left[Y\,|x,\,Z\right]\right] \\
&= \sum_z \mathbb{E}\left[Y\,|x,\,z\right] P(z) \\
\mathbb{E}\left[Y|do(X=1)\right] - \mathbb{E}\left[Y|do(X=0)\right] &= \mathbb{E}_Z\left[\mathbb{E}\left[Y\,|X=1,\,Z\right] - \mathbb{E}\left[Y\,|X=0,\,Z\right]\right]
\end{aligned}$$

Unlike the potential outcome theory, which do not know how to choose confounder Z. Using graphical causal models, we know how to choose a valid Z: we simply choose Z, so that it satisfies the back-door criterion. Then, under the assumptions encoded in the causal graph, *conditional exchangeability* provably holds; the causal effect is provably *identifiable*.

References

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 217–224, 2006.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *proceedings of the 21st national conference on Artificial intelligence-Volume 2*, pages 1219–1226, 2006.