Calculus of Interventions

Source: (a) Elements of Causal Inference, Peters et al., Chap 6.
(b) Causality, J. Pearl, 2009. $\xrightarrow{2}$ (aka "text")

Outline

(a) **Known:** SCM $C = (S, N_x)$, its associated Markov DAG $G = (V, E)$, and the Markov factorization of the associated joint dist. over $(x_1, \ldots, x_d)$.

(b) An intervention $X_j := \tilde{N}_j$ occurs.

(c) **Goal:** We want to compute the intervention distribution, but using only the observed distribution.

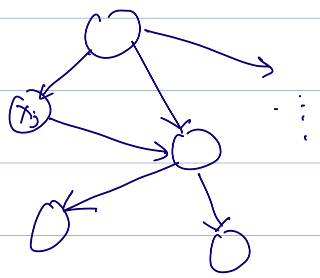
Markov Factorization with the Intervention DAG.

$\mathfrak{C} = (S, N)$ observational SCM with $G = (V, E)$

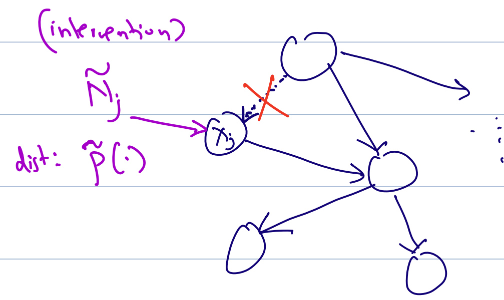
$\tilde{\mathfrak{C}} = (\tilde{S}, \tilde{N})$ the interventional SCM

For concreteness say $X_j := \tilde{N}_j$ is the intervention. Further let the dist of $\tilde{N}_j$ be $\tilde{P}(x_j)$.

indep. of $N_1, \ldots, N$ and over the alphabet of X_j.



$\mathcal{G}$ associated with $\mathfrak{C}$



$\tilde{\mathcal{G}}$ associated with $\tilde{\mathfrak{C}}$

$$p^{\mathfrak{C}}(x_1, x_2, \cdots, x_d) = \prod_i p^{\mathfrak{C}}(x_i \mid PA_i)$$

Observational dist. factorization

$$p^{\tilde{\mathfrak{C}}}(x_1, \cdots, x_d) = p^{\mathfrak{C};\, do(x_j := \tilde{N}_j)}(x_1, \cdots, x_d)$$

$$= \left(\prod_{i \neq j} p^{\mathfrak{C}}(x_i \mid PA_i) \right) \tilde{p}(x_j) \longrightarrow (1).$$

Intervention dist. factorization

none of the factors corresponding to non-intervened nodes change.

The above immediately follows from Theorem 3 (Notes 2), i.e., the Markov Property for a DAG.

Further, suppose $\tilde{N}_j \doteq a$ w.p. 1, i.e., an atomic intervention occurs.

Then, we can write (1) as:

$$p^{\mathfrak{C};\, do(X_j := a)}(x_1, \ldots, x_d) = \left(\prod_{i \neq j} p^{\mathfrak{C}}(x_i \mid PA_i) \right) \chi_{\{x_j = a\}} \longrightarrow \text{②}$$

where $\chi_{\{x_j=a\}} = \begin{cases} 1 & \text{if } x_j = a \\ 0 & \text{otherwise} \end{cases}$

(indicator function)

① and ② are referred to as the truncated factorization theorem (other names as well; see section 6.6 in text for details).

'Conditioning' = 'do' for nodes with no parents.

Suppose that X_1 is a node with no parents. Then,

$$p^{\mathfrak{C}}(x_2, \ldots, x_d \mid X_1 = a) = \frac{p^{\mathfrak{C}}(a, x_2, \ldots x_d)}{P^{\mathfrak{C}}(X_1 = a)}.$$

$$= \frac{\cancel{P^{\mathfrak{C}}(X_1 = a)}}{\cancel{P^{\mathfrak{C}}(X_1 = a)}} \cdot \prod_{i=2}^{d} p^{\mathfrak{C}}(x_i \mid \underbrace{PA_i}_{})$$

↳ $x_1 = a$, where ever this variable appears.

Now from (2),

$$P^{\mathfrak{C};do(X_1:=a)}(x_2,\dots,x_d) = \sum_{x_1} P^{\mathfrak{C};do(X_1:=a)}(x_1,x_2,\dots,x_d)$$

$$= \left(\prod_{i=2}^{d} P^{\mathfrak{C}}(x_i \mid \underbrace{PA_i}_{x_1=a})\right) \cdot \overbrace{1}^{\chi_{\{x_1=a\}}}$$

i.e., $\boxed{P^{\mathfrak{C}}(x_2,\dots,x_d \mid X_1=a) = P^{\mathfrak{C};do(X_1:=a)}(x_2,\dots,x_d).}$

for X_1 with no parents.

Example: (Kidney stones, see Notes 1, example 6.37 in text)

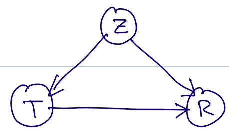
Kidney stone recovery data from 700 patients

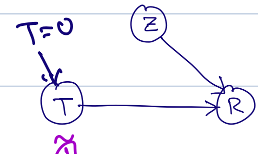
Successful Recovery Statistics

	Overall Success	Patients with Small stones	Patients with Large stones
Treatment a: Open Surgery	78% (273/350)	93% (81/87)	73% (192/263)
Treatment b: (small puncture surgery - percutaneous nephrolithotomy)	83% (289/350)	87% (234/270)	69% (55/80)

A causal, quantitative model of above:

$\mathfrak{C}$ (Causal model)





$\tilde{\mathfrak{C}}$

Z = size of stone $\in \begin{cases} 0 & \text{small} \\ 1 & \text{large} \end{cases}$

R = recovery / success of treatment

T = Treatment $\in \begin{cases} 0 & \text{Treatment } a \\ 1 & \text{Treatment } b \end{cases}$

$\in \begin{cases} 0 & \text{did not succeed} \\ 1 & \text{succeed} \end{cases}$

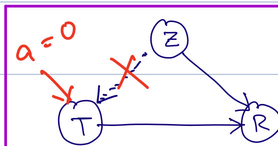
Treatment a ('0') $\longrightarrow$ open surgery

Treatment b ('1') $\longrightarrow$ small puncture

Goal: Compute $P_R^{\mathfrak{C};do(T:=0)}(1)$ and $P_R^{\mathfrak{C};do(T:=1)}(1)$

marginalize on recovery r.v.

$$P_R^{\mathfrak{C};do(T:=0)}(1) = P_{R,T,Z}^{\tilde{\mathfrak{C}}}(1,0,0)$$
$$+ P_{R,T,Z}^{\tilde{\mathfrak{C}}}(1,0,1)$$



$\tilde{\mathfrak{C}}$

Note that: $\quad p^{\mathfrak{C}}(r \mid PA_r) = p^{\tilde{\mathfrak{C}}}(r \mid PA_r)$

and $\quad p^{\mathfrak{C}}(z \mid PA_z) = p^{\tilde{\mathfrak{C}}}(z \mid PA_z)$.

Equivalently, from 2 :

$$\boxed{p^{\mathfrak{C};do(x_j:=a)}(x_1,\dots,x_d) = \left(\prod_{i\neq j} p^{\mathfrak{C}}(x_i \mid PA_i)\right) \chi_{\{x_j=a\}} \longrightarrow 2}$$

$$P_R^{\mathbb{C};do(T:=0)}(1) = P^{\mathbb{C}}_{R|T,Z}(1|0,0) \cdot P^{\mathbb{C}}_Z(0) \cdot \underbrace{1}_{\chi_{\{T=0\}}}$$
$$+ P^{\mathbb{C}}_{R|T,Z}(1|0,1) \cdot P^{\mathbb{C}}_Z(1) \cdot \underbrace{1}_{\chi_{\{T=0\}}}$$

$$\boxed{\text{These quantities can be estimated purely from observational data in Table.}}$$

$\therefore \quad P_R^{\mathbb{C};do(T:=0)}(1) \approx 0.832.$

Similarly, $\quad P_R^{\mathbb{C};do(T:=1)}(1) \approx 0.782.$

$\Longrightarrow$ open surgery better than small puncture

$\Longrightarrow$ no paradox !

Further, the average causal effect for binary treatments $= (0.832 - 0.782)$.

Again, as a reminder, the example above shows that intervening on T $\neq$ conditioning on T.

$$P^{\mathbb{C}}(R=1|T=0) = \frac{P^{\mathbb{C}}(R=1,Z=0,T=0) + P^{\mathbb{C}}(R=1,Z=1,T=0)}{P^{\mathbb{C}}(T=0)} = \frac{P^{\mathbb{C}}(R=1|Z=0,T=0)P^{\mathbb{C}}(Z=0|T=0)P^{\mathbb{C}}(T=0) + P^{\mathbb{C}}(R=1|Z=0,T=0)P^{\mathbb{C}}(Z=0|T=0)P^{\mathbb{C}}(T=0)}{P^{\mathbb{C}}(T=0)}$$

$$= P^{\mathbb{C}}_{T|R,Z}(1|0,0)P^{\mathbb{C}}_{Z|T}(0|0) + P^{\mathbb{C}}_{T|R,Z}(1|0,1)P^{\mathbb{C}}_{Z|T}(1|0)$$

Conditioning

Adjustment

Informally, "adjusting for a variable Z" or "Z is a valid adjustment set" means that we can compute intervention probabilities using conditional probabilities by conditioning on Z and the intervention variable, BUT unconditioning ONLY on the marginal of Z in the original, observational DAG / SCM.

This can be interpreted as an altered "Total Probability Theorem" that relates interventional distributions to observational distributions.

Defn: $\mathfrak{C}$ a causal model. Suppose that we intervene on X_k, i.e., $do(X_k := x_k)$. Consider any $X_j \notin PA_k$. Then, $\boldsymbol{Z} \subseteq \{X_1, \dots X_d\} \setminus \{X_j, X_k\}$ is called a valid adjustment set if

$$\boxed{P_{X_j}^{\mathfrak{C};do(X_k:=x_k)}(x_j) = \sum_{\underbrace{z}_{\text{vector of realizations in } \boldsymbol{Z}}} P^{\mathfrak{C}}(x_j \mid x_k, z)\, P^{\mathfrak{C}}(z)}$$

Example: In the Kidney Stones example, $\boldsymbol{Z} = \{Z\}$, i.e., we adjusted with Z (size of stone) to compute the average causal effect on R (recovery).

In general, we have seen that

$$P_{X_j}^{\mathfrak{C}; do(X_K := x_K)}(x_j) \neq P_{X_j | X_K}^{\mathfrak{C}}(x_j | x_K)$$

We refer to the above effect as 'confounding'. In general, we need to find a valid adjustment set to compute interventional prob. from observational data when X_K confounds the effect at X_j.

Defn (Confounding): The effect from X_K to X_j is confounded if:

$$P_{X_j}^{\mathfrak{C}; do(X_K := x_K)}(x_j) \neq P_{X_j | X_K}^{\mathfrak{C}}(x_j | x_K)$$

Getting around confounding: Determining Valid Adjustment Sets.

Roadmap on finding Adjustment Sets:

1. What is the property we are looking for? (Invariant Conditionals)

2. Is it always better to use bigger adjustments sets, i.e., adjust on as many variables/covariates as possible?
(No; Berkson's Paradox)

3. Characterizing Invariant Conditionals using d-separation

4. Putting things together: Adjustment theorem using graphs and d-separation.

---

1. (Invariant Conditionals) Z is a valid adjustment set if the following invariance holds across the observational and interventional SCMs:

$$\boxed{\begin{aligned} P^{\mathfrak{C};do(X_k=x_k)}(x_j \mid x_k, z) &= P^{\mathfrak{C}}(x_j \mid x_k, z) \\ P^{\mathfrak{C};do(X_k=x_k)}(z) &= P^{\mathfrak{C}}(z) \end{aligned}} \quad \circledast$$

i.e., the conditional on the target (above can be generalized to sets instead of scalars) is invariant when conditioned on the adjustment Z, and the marginals of Z remained unchanged.

Motivation for the above defn. of invariant conditionals:

Recall from the definition of " Z is a valid adjustment" that:

Below only true when Z is a valid adjustment

$$\boxed{P_{X_j}^{\mathbb{C};do(X_k:=x_k)}(x_j) = \sum_{\underset{\text{vector of realizations in } Z}{z}} P^{\mathbb{C}}(x_j \mid x_k, z)\, P^{\mathbb{C}}(z)}$$

Further, from the usual total probability theorem on the altered SCM $\mathbb{C};do(X_k:=x_k)$, we have

$$P_{X_j}^{\mathbb{C};do(X_k:=x_k)}(x_j) = \sum_{z} P^{\mathbb{C};do(X_k:=x_k)}(x_j \mid x_k, z)\, P^{\mathbb{C};do(X_k:=x_k)}(z)$$

(The above is always true for any Z)

The invariant conditional definition comes from term by term matching of these two expressions above.

---

2. Is it always better to have larger adjustment sets, i.e., if $Z = (z_1, \dots, z_l)$ is a valid adjustment set, is it true that (Z, Z') is also a

Valid adjustment? No!

Berkson's Paradox: (Example 6.30 in text)

The statement: "Why are handsome men such jerks?"
(example originally from Ellenberg 2014; .... Berkson 1946)

There are (implicitly) three variables here:

Is the man in a relationship? $R = \begin{cases} 0 & \text{no} \\ 1 & \text{yes.} \end{cases}$

Is the man handsome? $H = \begin{cases} 0 & \text{no} \\ 1 & \text{yes} \end{cases}$

Is the man friendly? $F = \begin{cases} 0 & \text{no} \\ 1 & \text{yes.} \end{cases}$

The underlying DAG is posited to be:

$$\begin{tikzcd} H \arrow[dr] & & F \arrow[dl] \\ & R & \end{tikzcd}$$

$H := N_H \sim \text{Bernoulli}(0.5)$

$F := N_F \sim \text{Bernoulli}(0.5)$

$R := \min(H, F) \oplus N_R$

$\hookrightarrow \sim \text{Bernoulli}(0.1)$

i.e., in words: If a man is both friendly and handsome, he is more likely to be in a relationship than not.

In this case: $H \perp\!\!\!\perp F$, but $H \not\perp\!\!\!\perp F \mid R$.

i.e., $Z = \emptyset$ is a valid adjustment set if we intervened on H, and observed the effect on F

i.e., $P_F^{\mathfrak{C}; do(H:=1)}(1) = P_F^{\mathfrak{C}}(1 \mid H=1)$

$= P_F^{\mathfrak{C}}(1).$

BUT if $Z = R$, then Z "anticorrelates" H and F, and the adjustment is no longer true!

---

3. Characterizing valid adjustments using d-separation on an augmented graph.

Recall pp. 5 in Notes 4 (copied below as an inset). Let us consider the augmented model $\mathfrak{C}^*$ that encodes interventions through the new variables $\{I_j\}$. For simplicity, let us assume that we intervene only on node k, and with the associated variable I_k, with

$$P(I_k=0) = P(I_k=1) = 0.5$$

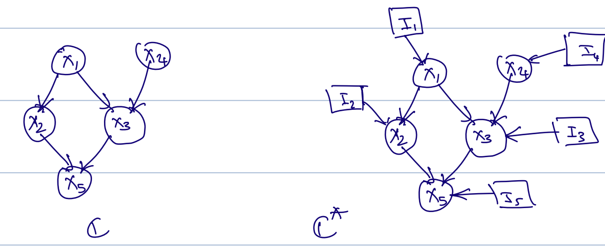
Further, for simplicity, let us assume that when intervened, $X_k := x_k$, i.e.,

$$X_k := \begin{cases} f_k(PA_k, N_k) & \text{if } I_k = 0 \\ x_k & \text{if } I_k = 1 \end{cases}$$

Alternative way to formalize Interventions using Intervention Variables

SCM $\mathfrak{C} = (S, P_N)$ over $(X_1, \cdots, X_d)$

For each node X_j, associate a new, additional parent node I_j, $j = 1, 2, \dots d$ called intervention variables.



where $X_j = \begin{cases} f_j(PA_j, N_j) & \text{if } I_j = \text{idle} \\ I_j & \text{otherwise.} \end{cases}$

I_j (when active) encodes the intervention, i.e., I_j can take values $\{x_{i_1}, \dots x_{i_d}\}$, with an intervention pmf.

Then, $P_Y^{\mathfrak{C}; do(X_j := x_{i_j})}(\cdot) = P_{Y \mid I_j = x_{i_j}}^{\mathfrak{C}^*}(\cdot)$

measure on some target variable Y.

i.e., interventions can be computed using the usual conditional dist. on the augmented $\mathfrak{C}^*$ model

Then, $p^{\mathfrak{C}^*}(x_1, \cdots, x_d \mid I_k = 0) = p^{\mathfrak{C}}(x_1, \cdots, x_d)$

and

$$p^{\mathfrak{C}^*}(x_1, \dots, x_d \mid I_k = 1) = p^{\mathfrak{C}; do(X_k = x_k)}(x_1, \dots, x_d)$$

Theorem 3, Notes 2

Recall that from Global Markov Property for $\mathfrak{C}^*$, d-separation $\Longrightarrow$ conditional independence.

i.e., with $Y \subseteq \{X_1, \dots, X_d\}$, $W \subseteq \{X_1, \dots, X_d\}$
$Y \cap W = \emptyset$:

$$Y \perp\!\!\!\perp_{\mathcal{G}^*} I \mid W \Rightarrow p^{\mathcal{C}^*}(\mathbf{y} \mid \boldsymbol{\omega}, I=0) = p^{\mathcal{C}^*}(\mathbf{y} \mid \boldsymbol{\omega}, I=1).$$

∴ From above,

$$p^{\mathcal{C}}(\mathbf{y} \mid \boldsymbol{\omega}) = p^{\mathcal{C}^*}(\mathbf{y} \mid \boldsymbol{\omega}, I=0) = p^{\mathcal{C}^*}(\mathbf{y} \mid \boldsymbol{\omega}, I=1) = p^{\mathcal{C}, do(X_k := x_k)}(\mathbf{y} \mid \boldsymbol{\omega})$$

$$\Rightarrow p^{\mathcal{C}}(\mathbf{y} \mid \boldsymbol{\omega}) = p^{\mathcal{C}; do(X_k := x_k)}(\mathbf{y} \mid \boldsymbol{\omega}) \rightarrow (**)$$

∴ in summary, sufficient conditions to characterize a valid adjustment set are (using (*) and (**))

$$\boxed{\begin{aligned} &p^{\mathcal{C}; do(X_k = x_k)}(x_j \mid x_k, z) = p^{\mathcal{C}}(x_j \mid x_k, z) \\ &p^{\mathcal{C}; do(X_k = x_k)}(z) = p^{\mathcal{C}}(z) \end{aligned}} \Bigg\} \rightarrow (*)$$

$$\boxed{\underbrace{X_j}_{\text{target}} \perp\!\!\!\perp_{\mathcal{G}^*} I \mid (\underbrace{X_k}_{\text{node that we are intervening on}}, Z) \; (a) \quad \text{and} \quad Z \perp\!\!\!\perp_{\mathcal{G}^*} I \; (b)}$$

---

(4) Finally, the main theorem on valid adjustment sets (Prop. 6.41 in text). There are three

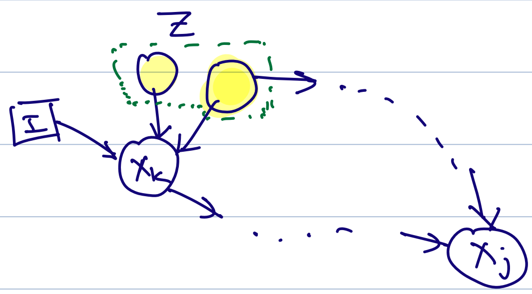
sufficient conditions. I am listing only two; please see text for the third.

Theorem (Valid Adjustment Sets) SCM $\mathfrak{C}$, with:

X_k the node where intervention occurs

X_j the target variable

ⓐ **Parental adjustment:** $Z = PA_k$ a valid adjustment for (X_k, X_j).



address both ⓐ and ⓑ

$X_j \perp\!\!\!\perp_{G^*} I \mid X_k, Z$

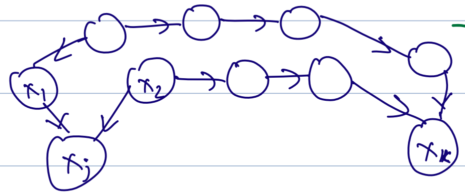
$Z \perp\!\!\!\perp_{G^*} I$

ⓑ **Backdoor criterion:** $Z \subseteq \{X_1, \dots X_d\} \setminus \{X_j, X_k\}$ s.t.

(i) Z contains NO descendants of X_k → addresses ⓑ

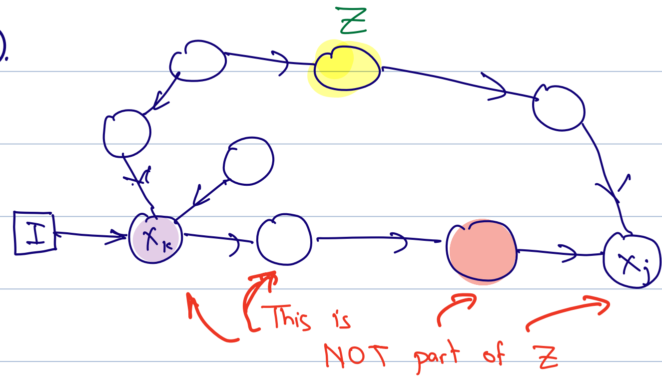
(ii) Z blocks all paths from X_k to X_j that passes through the parents of X_k (i.e., "backdoor" paths). → addresses ⓐ

note that X_k itself blocks all direct paths from $I \to X_j$



these are backdoor paths from X_j to X_k

example: (with G^* graph)



ⓐ $X_j \perp\!\!\!\perp_{G^*} I \mid X_k, Z$: Above, (X_k, Z) blocks all paths from I to X_j.

Note that conditioning on X_k alone OPENS the backdoor path from I to X_j because X_k is a collider. Therefore, we need to additionally add Z to block this backdoor path.

ⓑ $Z \perp\!\!\!\perp_{G^*} I$: This is true because we are not conditioning here on X_k or any of its descendents. Note that if descendants of X_k were included, then X_k (and its descendants) is a collider, and unblocks the path from I to Z.

Beyond Adjustments — do-calculus.

In general, we would like to compute interventional dist. from observational distributions. Adjustment allowed one approach to determine invariant conditionals. More generally, we care about identifiability of an interventional distribution, meaning:

Can $P_Y^{\mathfrak{C}; do(X:=x)}(y)$ be computed purely from observational distribution?

do-calculus provides a set of rules that allows manipulation of a conditional dist. in the intervention SCM. These rules provide other way (beyond adjustment) for computing intervention dist.

Refs: do-calculus is at the core of Pearl's framework for causality, see Pearl's book, 2009. See also text (by Peters et. al.) sec. 6.7 for an abridged discussion. The discussion here follows the text by Peters et al.

Setup: SCM $\mathfrak{C}$, Graph $\mathcal{G}$

$X \to$ node(s) on which intervention being done

$Y \to$ target variable(s)

X, Y, Z, W disjoint (sets) of nodes.

ⓐ (Insertion/Deletions of Observations)

Suppose $Y \perp\!\!\!\perp_{\tilde{\mathcal{G}}} Z \mid X, W$ where $\tilde{\mathcal{G}}$ is the DAG with (incoming edges to X) have been removed.
Then:

$$P^{\mathfrak{C}; do(X:=x)}(y \mid z, w) = P^{\mathfrak{C}; do(X:=x)}(y \mid w).$$

ⓑ (Action/Observation Exchange)

Suppose $Y \perp\!\!\!\perp_{\hat{\mathcal{G}}} Z \mid X, W$ where $\hat{\mathcal{G}}$ is the DAG where (incoming edges to X) and (outgoing edges from Z) are removed. Then:

$$P^{\mathfrak{C}; do(X:=x, Z:=z)}(y \mid w) = P^{\mathfrak{C}; do(X:=x)}(y \mid z, w).$$

ⓒ (Insertion | Deletion of Actions)

Suppose $Y \perp\!\!\!\perp_{\mathcal{G}'} Z \mid X, W$ where $\mathcal{G}'$ is the DAG where (incoming edges to X) and (incoming edges to $Z(W)$) have been removed.

$$Z(W) = \left\{ X_\ell \in Z \text{ s.t. } X_\ell \text{ not an ancestor of } W \text{ in } \tilde{\mathcal{G}} \right\}$$

Then:

$$P^{\mathfrak{C};do(X:=x, Z:=z)}(y \mid w) = P^{\mathfrak{C};do(X:=x)}(y \mid w).$$

Theorem (6.45 in text):

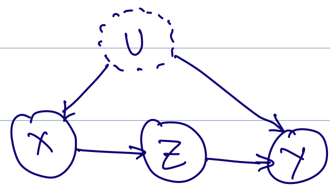
The rules above are complete, i.e., all identifiable intervention distributions can be computed using these three rules.

In addition, $\exists$ an algorithm by Tian 2002 that can use these rules to find all identifiable intervention distributions

---

Finally to conclude, we state a result has been proved by applying the rules above.

Prop (Front door Criterion) Let $\mathfrak{C}$ be an SCM with DAG $\mathcal{G}$ given by: (assume $p^{\mathfrak{C}}(x,z)>0$)



Note: dotted circle means hidden/latent variable

(Note that since U is not observed, we cannot use backdoor adjustment to study interventions on X and effect on Y. In fact, there is no valid adjustment set in this case.).

Then:
$$p^{\mathfrak{C};do(X:=x)}(y) = \sum_z p^{\mathfrak{C}}(z|x) \cdot \left(\sum_{\tilde{x}} p^{\mathfrak{C}}(y|\tilde{x},z)\, p^{\mathfrak{C}}(\tilde{x}) \right)$$