Causality Basics

(Based on Peters et al., 2017,
“Elements of Causal Inference: Foundations and Learning Algorithms”)

Chang Liu

2020.12.16

Content

- Causality and Structural Causal Models.
- Causal Inference.
- Causal Discovery.

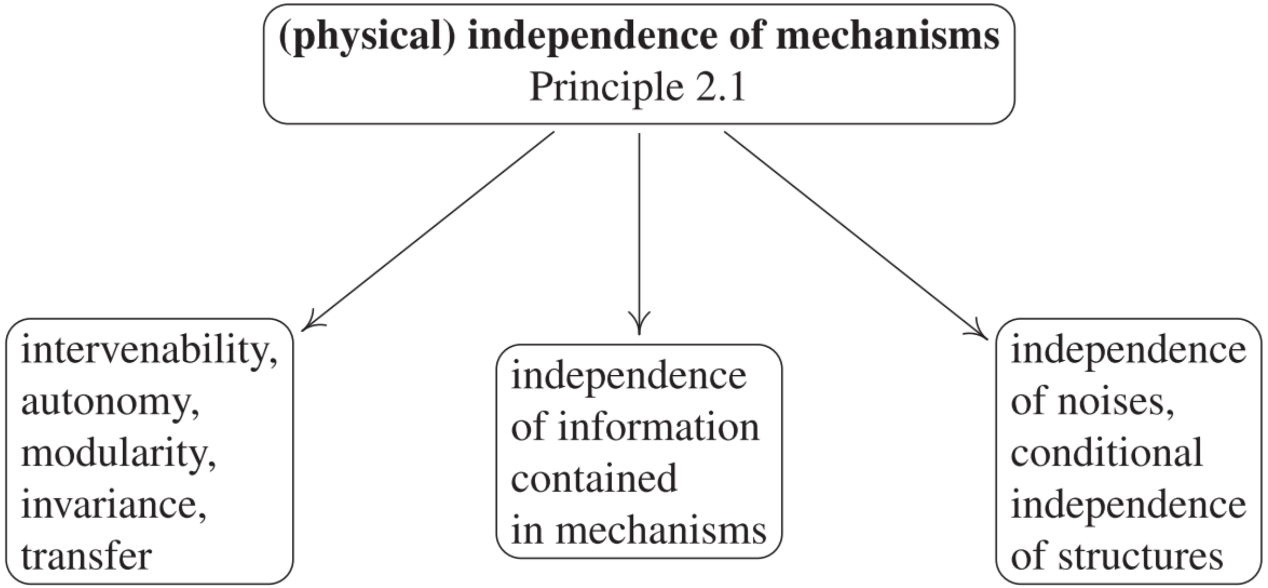
Causality

- Formal definition of causality:
  "two variables have a causal relation, if *intervening* the cause may change the effect, but not *vice versa*" [Pearl, 2009; Peters et al., 2017].
  - Intervention: change the value of a variable by leveraging mechanisms and changing variables out of the considered system.
- Example: for the Altitude and average Temperature of a city, $A \to T$.
  - Running a huge heater (intv. T) does not lower A.
  - Raising the city by a huge elevator (intv. A) lowers T.
- Causality contains more information than observation (static data).
  - E.g., both $p(A)p(T|A)$ ($A \to T$) and $p(T)p(A|T)$ ($T \to A$) can describe the observed relation $p(A,T)$.

Causality

The Independent Mechanism Principle.

- **Principle 2.1 (Independent mechanisms)**
  The causal generative process of a system's 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.

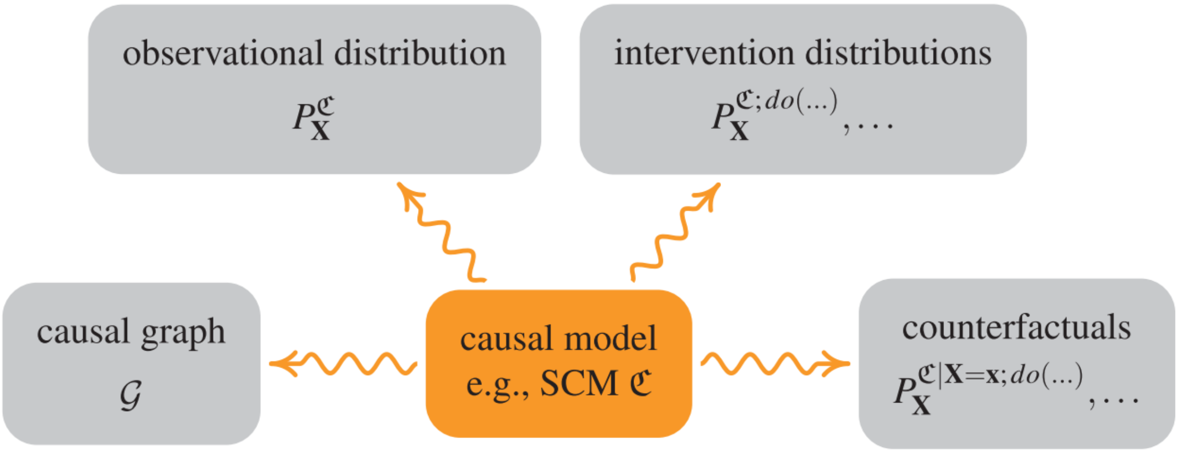


- It is possible to conduct a *localized* intervention.
- We can change one mechanism without affecting the others.

Structural Causal Models

For describing causal relations.

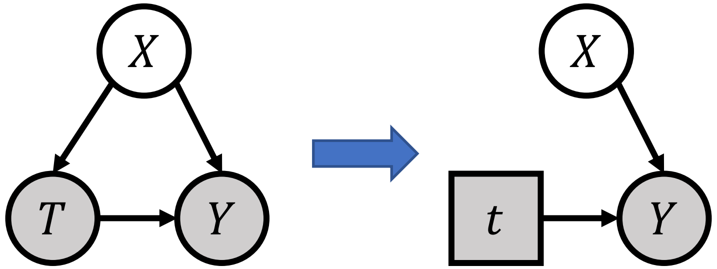
- **Definition 6.2 (Structural causal models)** An SCM $\mathfrak{C} := (\mathbf{S}, P_{\mathbf{N}})$ consists of a collection $\mathbf{S}$ of d **(structural) assignments**, $X_j := f_j(\mathrm{pa}_j, N_j), j = 1, \ldots, d$ where pa_j are called parents of X_j, and a joint distr. $P_{\mathbf{N}} = P_{N_1,\ldots,N_d}$ over the noise variables, which we require to be jointly independent.
  - The joint indep. requirement on $\mathbf{N}$ comes from the Independent Mechanism Principle (the right block).
- **Causal graph** $\mathcal{G}$: the Directed Acyclic Graph constructed based on the assignments.
- **Proposition 6.3 (Entailed distributions)** An SCM $\mathfrak{C}$ defines a unique distribution over the variables $\mathbf{X} = (X_1, \ldots, X_d)$ such that $X_j = f_j(\mathrm{pa}_j, N_j), j = 1, \ldots, d$, in distribution, called the entailed distribution $P_{\mathbf{X}}^{\mathfrak{C}}$ and sometimes write $P_{\mathbf{X}}$.
- Implications of an SCM:



Structural Causal Models

Describing interventions using SCMs.

- "What if X was *set* to x?""What if patients took the treatment?"
- **Definition 6.8 (Intervention distribution)** Consider an SCM $\mathfrak{C} := (\mathbf{S}, P_{\mathbf{N}})$ and its entailed distribution $P_{\mathbf{X}}^{\mathfrak{C}}$. We *replace* one (or several) of the structural assignments to obtain a new SCM $\tilde{\mathfrak{C}}$. Assume that we replace the assignment for X_k by: $X_k := \tilde{f}\left(\widetilde{\mathrm{pa}}_k, \tilde{N}_k\right)$, where all the new and old noises are required to be jointly independent.
  - **Atomic/ideal**, **structural** [Eberhardt and Scheines, 2007]/**surgical** [Pearl, 2009]/**independent /deterministic** [Korb et al., 2004] **intervention**:
    when $\tilde{f}\left(\widetilde{\mathrm{pa}}_k, \tilde{N}_k\right)$ puts a point mass on a value a. Simply write $P_{\mathbf{X}}^{\mathfrak{C};do(X_k:=a)}$.
  - **Soft intervention**: intervene with a distribution.



Structural Causal Models

Describing interventions using SCMs.

- **Definition 6.12** (**Total causal effect**) Given an SCM $\mathfrak{C}$, there is a total causal effect from X to Y, if X and Y are *dependent* in $P_{\mathbf{X}}^{\mathfrak{C};do(X:=\tilde{N}_X)}$ for some r.v. $\tilde{N}_X$.
- **Proposition 6.13 (Total causal effects)** Equivalent statements given an SCM $\mathfrak{C}$:
  1. There is a total causal effect from X to Y.
  2. There are $x^{(1)}$ and $x^{(2)}$ such that $P_Y^{\mathfrak{C};do\left(X:=x^{(1)}\right)} \neq P_Y^{\mathfrak{C};do\left(X:=x^{(2)}\right)}$.
  3. There is $x^{(1)}$ such that $P_Y^{\mathfrak{C};do\left(X:=x^{(1)}\right)} \neq P_Y^{\mathfrak{C}}$.
  4. X and Y are dependent in $P_{X,Y}^{\mathfrak{C};do(X:=\tilde{N}_X)}$ for any $\tilde{N}_X$ with full support.
- **Proposition 6.14 (Graphical criteria for total causal effects)** Consider SCM $\mathfrak{C}$ with corresponding graph $\mathcal{G}$.
  - If there is no directed path from X to Y, then there is no total causal effect.
  - Sometimes there is a directed path but no total causal effect.

Structural Causal Models

Describing counterfactuals using SCMs.

- "What if *I* took the treatment?"
  "What would the outcome be had I acted differently *at that moment/situation*?"
- Intervention {in a given situation/context} / {for a given individual/unit}.
  Intervention with a conditional.
- **Definition 6.17 (Counterfactuals)** Consider SCM $\mathfrak{C} := (\mathbf{S}, P_{\mathbf{N}})$. Given some observations $\mathbf{x}$, define a *counterfactual* SCM by replacing the distribution of noise variables:
  $\mathfrak{C}_{\mathbf{X}=\mathbf{x}} := \left(\mathbf{S}, P_{\mathbf{N}}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}\right)$, where $P_{\mathbf{N}}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}} := P_{\mathbf{N}|\mathbf{X}=\mathbf{x}}$ (need not be jointly independent anymore).
  - "The probability that Z would have been 7 (say) had Y been set to 2 for observation $\mathbf{x}$":

2. the observed data $\mathbf{X} = \mathbf{x}$

1. the SCM $\mathfrak{C}$ we start with

3. the intervention $do\,(Y := 2)$

$$P_Z^{\mathfrak{C}|\mathbf{X}=\mathbf{x};\,do(Y:=2)}$$

4. the variable Z we are interested in

Structural Causal Models

Observational implications (SCM as a BayesNet); causality and correlation.

- **Definition 6.1 (d-separation)**. A path P is *d-separated* by a set of nodes $\boldsymbol{S}$, iff. P either has an emitter ($\to X \to$ or $\leftarrow X \to$) in $\boldsymbol{S}$, or a collider ($\to X \leftarrow$) that is not in $\boldsymbol{S}$ nor are its descendants.

- **Definition 6.21 (Markov property)** *Given a DAG $\mathcal{G}$ and a joint distribution $P_{\mathbf{X}}$, this distribution is said to satisfy*
  (i) the ***global Markov property*** *with respect to the DAG $\mathcal{G}$ if*

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \,|\, \mathbf{C} \;\Rightarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

  for all disjoint vertex sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (the symbol $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation — see Definition 6.1),
  (ii) the ***local Markov property*** *with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and*
  (iii) the ***Markov factorization property*** *with respect to the DAG $\mathcal{G}$ if*

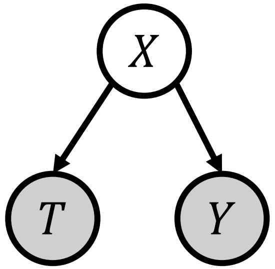
$$p(\mathbf{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j \,|\, \mathbf{pa}_j^{\mathcal{G}}).$$

- **Theorem 6.22 (Equivalence of Markov properties)**. When $P_{\mathbf{X}}$ has a density.
- (iii) is what we mean by using a DAG/BeyesNet to define a joint distribution (i.e., the entailed distribution by the SCM). Its equivalence to (i) means that we can read off conditional independence assertions of this joint distribution from the graph.

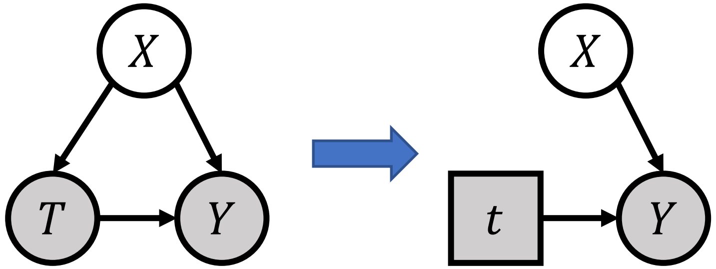
Structural Causal Models

Observational implications (SCM as a BayesNet); causality and correlation.

- Confounding bias.
  - T and Y are correlated if X is not given.
  - E.g., T= chocolate consumption, Y= #Nobel prices, X= development status.

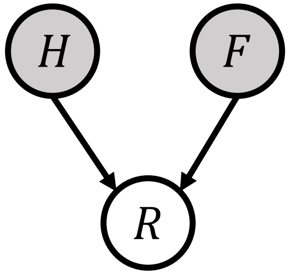


- Simpson's paradox.
  - T: treatment. Y: recovery. X: socio-economic status.
  - Is the treatment effective for recovery?
    $p(Y=1|T=1) > p(Y=1|T=0)$, but
    $p^{do(T=1)}(Y=1) < p^{do(T=0)}(Y=1)$.
  - $p(Y|t) = \sum_x p(Y|x,t)p(x|t)$:
    conditioning on $T=1$ infers a good X thus a higher probability of recovery.



Structural Causal Models

Observational implications (SCM as a BayesNet); causality and correlation.



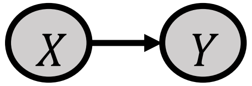
- Selection bias.
  - H and F are correlated if secretly/unconsciously conditioned on R.
- Berkson's paradox.
  - "Why are handsome men such jerks?" [Ellenberg, 2014]
  - H= handsome, F= friendly, R= in a relation.
  - Single women often date men with $R = 0$. Such men tend to be either not H or not R.
- (spurious correlations).

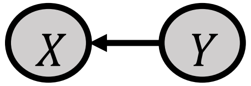
Structural Causal Models

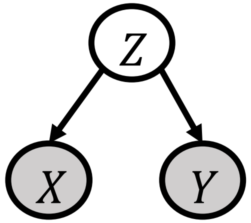
Levels of model according to the prediction ability:

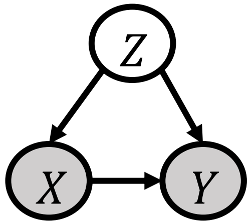
$$\underbrace{\underbrace{\underbrace{P_X}_{\substack{\textbf{probabilistic/observational model} \\ \text{(machine learning models)}}} + \text{intv. distrs.}}_{\textbf{interventional model}\ \text{(causal graphical models)}} + \text{counterf. statements}}_{\textbf{counterfactual model}\ \text{(SCMs)}}$$

- Interventional model has the utility of a probabilistic/observational model $p(X, Y)$:









$p(X,Y) =$	$p(X)p(Y\|X)$	$p(Y)p(X\|Y)$	$\int p(z)p(X\|z)p(Y\|z)\,\mathrm{dz}$	$\int p(z)p(X\|z)p(Y\|X,z)\,\mathrm{dz}$

  but cannot determine the interventional distribution $p^{do(X=x)}(Y)$:

$p^{do(X=x)}(Y) =$	$p(Y\|x)$	$p(Y)$	$\int p(z)p(Y\|z)\,\mathrm{dz} = p(Y)$	$\int p(z)p(Y\|x,z)\,\mathrm{dz}$

- **Example 6.19**: Two SCMs that induce the same graph, observational distributions, and intervention distributions, could entail different counterfactual statements ("probabilistically and interventionally equivalent" but not "counterfactually equivalent").

Structural Causal Models

Levels of model according to the prediction ability:

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

CAUSAL HIERARCHY THEOREM

[Bareinboim, Correa, Ibeling, Icard, 2020]

Given that an SCM M → PCH, we can show the following:

Theorem (CHT). With respect to Lebesgue measure over (a suitable encoding of L_3-equivalence classes of) SCMs, the subset in which any PCH ‘collapse’ is measure zero.

Informally, for almost any SCM (i.e., almost any possible environment), the PCH does not collapse, i.e., the layers of the hierarchy remains distinct.

Corollary. To answer question at Layer i (about a certain interaction), one needs knowledge at layer i or higher.

Causal Inference

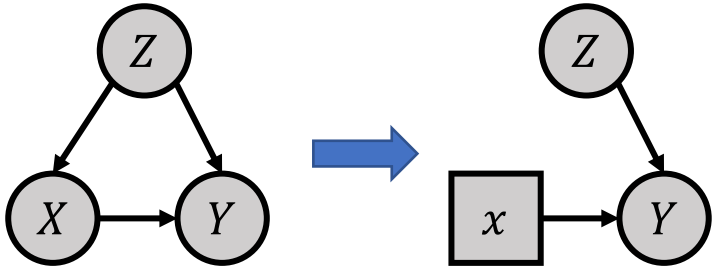
- Given an SCM, infer the distribution/outcome of a variable, under an intervention or unit-level intervention (counterfactual).
  - Estimate causal effect:
    - Average treatment effect: $\mathbb{E}^{do(T=1)}[Y] - \mathbb{E}^{do(T=0)}[Y]$.
    - Individual treatment effect: $\mathbb{E}^{do(T=1)}[Y|x] - \mathbb{E}^{do(T=0)}[Y|x]$.
- Estimate the interventional distribution is the key task (counterfactual is the conditional in an intervened SCM).

Causal Inference

- **Truncated factorization** [Pearl, 1993] / **G-computation formula** [Robins, 1986] / **manipulation theorem** [Spirtes et al., 2000]:
  $p^{\mathfrak{C};do(X_k:=\tilde{N}_k)}(x_1, \ldots, x_d) = \tilde{p}(x_k) \prod_{j \neq k} p^{\mathfrak{C}}(x_j | x_{\mathrm{pa}_j})$, where $\tilde{p}$ is the density of $\widetilde{N}_k$.
  - Almost evident from the interventional SCM.
  - Relies on the autonomy/modularity of causality (principle of independence of mechanisms):
    one causal mechanism does not influence or inform other causal mechanisms.

Causal Inference

- **Definition 6.38 (Valid adjustment set, v.a.s)** Consider an SCM $\mathfrak{C}$ over nodes $\mathbf{V}$. Let $X, Y \in \mathbf{V}$ with $Y \notin \mathrm{pa}_X$. We call a set $Z \subseteq \mathbf{V} \backslash \{X, Y\}$ a v.a.s for (X, Y) if $p^{\mathfrak{C};do(X:=x)}(y) = \sum_{\mathbf{z}} p^{\mathfrak{C}}(y|x, \mathbf{z}) p^{\mathfrak{C}}(\mathbf{z})$.
  - Importance: we can then use the *observations* of variables in a v.a.s to estimate an *intervention* distribution.
  - By the previous result, pa_X is a v.a.s.



  - Are there any other v.a.s's?

Causal Inference

- **Proposition 6.41** (Valid adjustment sets) The following $\mathbf{Z}$'s are v.a.s's for (X, Y) where $X, Y \in \mathbf{X}$ and $Y \notin \mathrm{pa}_X$:
  1. "**Parent adjustment**": $\mathbf{Z} \coloneqq \mathrm{pa}_X$;
  2. "**Backdoor criterion**" (Pearl's admissible/sufficient set): Any $\mathbf{Z} \subseteq \mathbf{X} \backslash \{X, Y\}$ with:
  - $\mathbf{Z}$ contains no descendant of X, and
  - $\mathbf{Z}$ blocks all paths from X to Y entering X through the backdoor ($X \leftarrow \cdots Y$);
  3. "**Toward necessity**" (characterize all v.a.s) [Shpitser et al., 2010; Perkovic et al., 2015]: Any $\mathbf{Z} \subseteq \mathbf{X} \backslash \{X, Y\}$ with:
  - $\mathbf{Z}$ contains no nodes (X excluded) on a directed path from X to Y nor their descendants, and
  - $\mathbf{Z}$ blocks all non-directed paths from X to Y.

Causal Inference

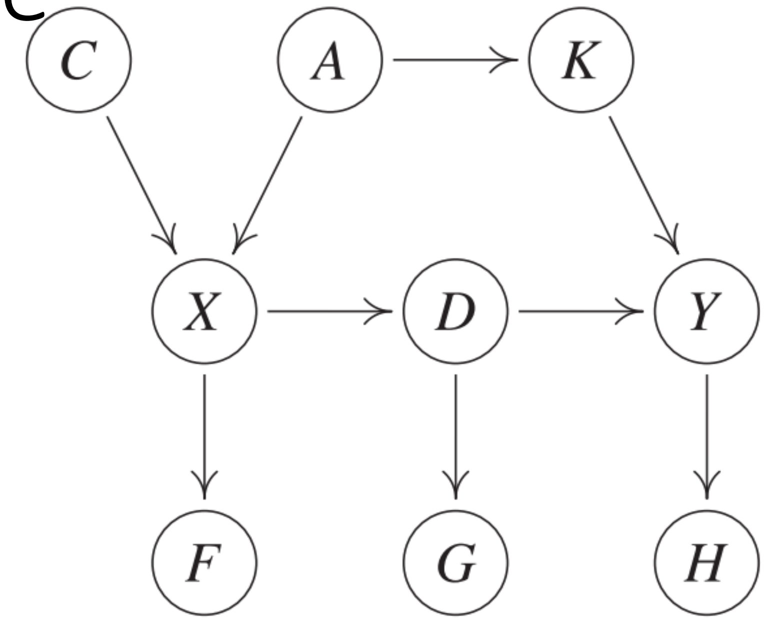


Figure 6.5: Only the path $X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y. The set $\mathbf{Z} = \{K\}$ satisfies the backdoor criterion (see Proposition 6.41 (ii)); but $\mathbf{Z} = \{F, C, K\}$ is also a valid adjustment set for (X, Y); see Proposition 6.41 (iii).

2. "**Backdoor criterion**" (Pearl's admissible/sufficient set): Any $\mathbf{Z} \subseteq \mathbf{X} \backslash \{X, Y\}$ with:
- $\mathbf{Z}$ contains no descendant of X, and
- $\mathbf{Z}$ blocks all paths from X to Y entering X through the backdoor ($X \leftarrow \cdots Y$);

3. "**Toward necessity**" (characterize all v.a.s): Any $\mathbf{Z} \subseteq \mathbf{X} \backslash \{X, Y\}$ with:
- $\mathbf{Z}$ contains no nodes (X excluded) on a directed path from X to Y nor their descendants, and
- $\mathbf{Z}$ blocks all non-directed paths from X to Y.

Causal Inference

Do-calculus: alternative way to estimate an *interventional* distribution from *observations*:

- The three rules of **do-calculus**: consider disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ from DAG $\mathcal{G}$.
  1. "Insertion/deletion of observations": $p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x})}(\mathbf{y}|\mathbf{z},\mathbf{w}) = p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x})}(\mathbf{y}|\mathbf{w})$, if $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ have been removed.
  2. "Action-observation exchange": $p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x},\mathbf{Z}:=\mathbf{z})}(\mathbf{y}|\mathbf{w}) = p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x})}(\mathbf{y}|\mathbf{z},\mathbf{w})$, if $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ and outgoing edges from $\mathbf{Z}$ have been removed.
  3. "Insertion/deletion of actions": $p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x},\mathbf{Z}:=\mathbf{z})}(\mathbf{y}|\mathbf{w}) = p^{\mathfrak{C};do(\mathbf{X}:=\mathbf{x})}(\mathbf{y}|\mathbf{w})$, if $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ and $\mathbf{Z}(\mathbf{W})$ have been removed. Here, $\mathbf{Z}(\mathbf{W})$ is the subset of nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$ in a graph that is obtained from $\mathcal{G}$ after removing all edges into $\mathbf{X}$.

Causal Inference

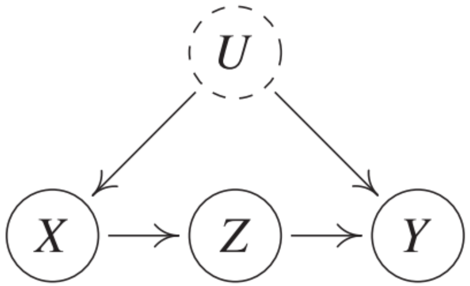
Do-calculus: alternative way to estimate an *interventional* distribution from *observations*:

- **Theorem 6.45** (Do-calculus).
  1. The three rules are complete: all identifiable intv. distrs. can be computed by an iterative application of these three rules [Huang and Valtorta, 2006, Shpitser and Pearl, 2006] (more general than v.a.s).
  2. There is an algorithm [Tian, 2002] that is guaranteed [Huang and Valtorta, 2006, Shpitser and Pearl, 2006] to find *all* identifiable intv. distrs.
  3. There is a nec. & suff. graphical criterion for identifiability of intv. distrs. [Shpitser and Pearl, 2006, Corollary 3], based on so-called hedges [see also Huang and Valtorta, 2006].

Causal Inference

Do-calculus: alternative way to estimate an *interventional* distribution from *observations*:

- **Example 6.46** (Front-door adjustment).



  - There is no v.a.s if we do not observe U.
  - But $p^{\mathfrak{C};do(X:=x)}(y)$ is identifiable using the do-calculus if $p^{\mathfrak{C}}(z|x) > 0$: $p^{\mathfrak{C};do(X:=x)}(y) = \sum_z p^{\mathfrak{C}}(z|x) \sum_{x'} p^{\mathfrak{C}}(y|x', z) p^{\mathfrak{C}}(x')$.
    - Observing Z in addition to X and Y reveals causal information nicely.
    - Causal relations can be explored by observing the “channel” Z that carries the “signal” from X to Y.

Causal Discovery

Learn a causal model (causal graph of an SCM) from data.

- In most cases, only observational data is available: Observational causal discovery.

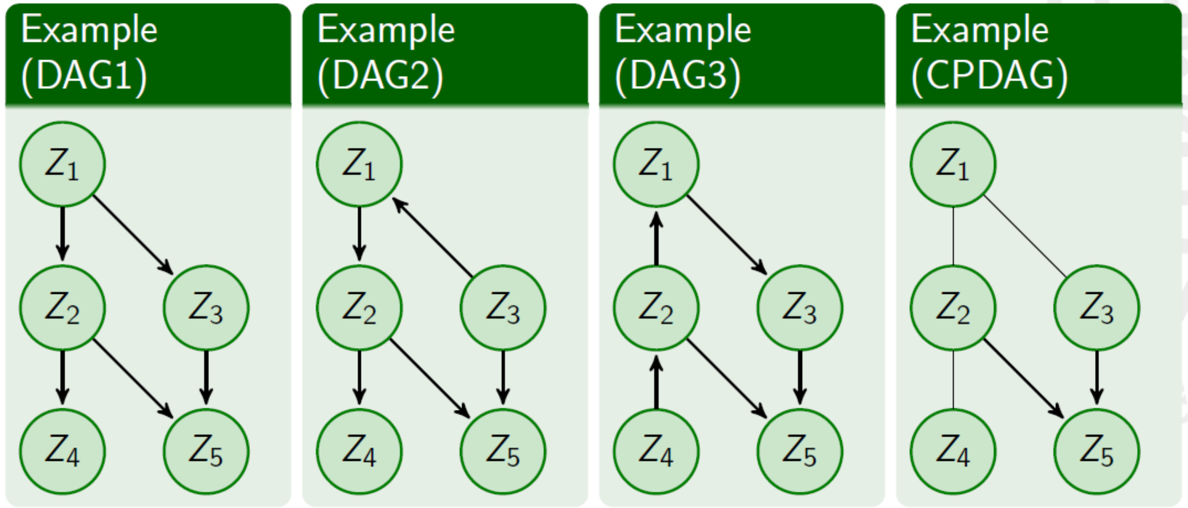
Two main-stream methods:

- Conditional Independence (CI) / constraint based methods (PC algorithm).
- Directional methods (additive noise models).

Causal Discovery

CI-based methods.

- Assumption:
  - $P_{\mathbf{X}}$ is Markovian (makes CI assertions as the graph does, in one of the three equiv. forms) and faithful (makes no more CI assertions than the graph does).
  - The Markov equivalence class is then identifiable (Lemma 7.2; Spirtes et al., 2000).
    - **Definition 6.24 (Markov equivalence of graphs)**: Two DAGs are said to be Markov. equiv., if they have the same set of distributions Markovian w.r.t them. Equivalently, they have the same set of d-separation.
    - **Lemma 6.25** (**Graphical criteria for Markov equivalence**; Verma and Pearl [1991]; Frydenberg [1990]) Two DAGs are Markov equiv., iff. they have the same skeleton and the same immoralities (v-structures).



Causal Discovery

CI-based methods.

- Instances:
  - Inductive causation (IC) algorithm [Pearl, 2009].
  - SGS algorithm (Spirtes, Glymour, Scheines) [Spirtes et al., 2000].
  - PC algorithm (Peter Spirtes & Clark Glymour) [Spirtes et al., 2000].

Causal Discovery

CI-based methods.

- Algorithm:
  - Estimation of Skeleton:
    - **Lemma 7.8** [Verma and Pearl, 1991, Lemma 1] (i) Two nodes X, Y in a DAG are adjacent, iff. they cannot be d-separated by any subset $S \subseteq \mathbf{X} \backslash \{X, Y\}$. (ii) If two nodes X, Y in a DAG are not adjacent, then they are d-separated by either pa_X or pa_Y.
      - IC/SGS: For each pair of nodes (X, Y), search through all such S's. X, Y are adjacent iff. they are cond. dep. given any one of such S's (i).
      - PC: efficient search for S. Start with a fully connected undirected graph, and traverse over such S's in an order of increasing size. For size k, one only has to go through S's that are subsets either of the neighbors of X or of the neighbors of Y (inverse negative of (ii)).

Causal Discovery

CI-based methods.

- Algorithm:
  - Orientation of Edges:
    - For a structure $X - Z - Y$ with no edge betw. X, Y in the skeleton, let A be a set that d-separates X and Y. Then the structure can be oriented as $X \to Z \leftarrow Y$, iff. $Z \notin A$.
    - More edges can be oriented in order to, e.g., avoid cycles.
    - Meek's orientation rules [Meek, 1995]: a complete set of orientation rules.

Causal Discovery

Additive Noise Models (ANM).

- Philosophy:

  Define a family of “particularly natural” conditionals, such that if a conditional can be described using this family, then the conditional in the opposite direction cannot be covered by this family.

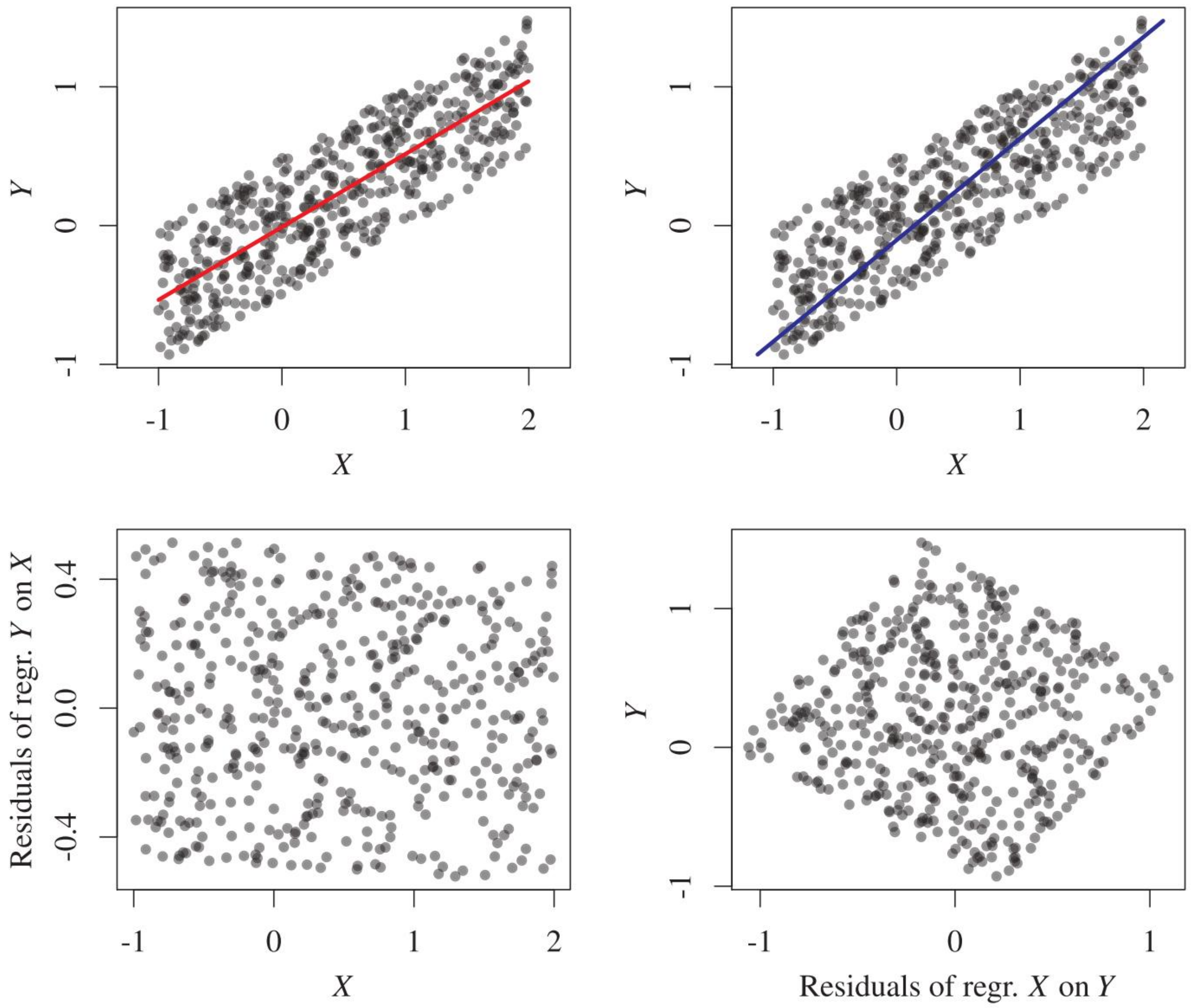
  - Determines a causal direction even for two variables with observational data.

- **Definition 7.3 (ANMs)** We call an SCM $\mathfrak{C}$ an ANM if the structural assignments are of the form $X_j \coloneqq f_j(\mathrm{pa}_j) + N_j, j = 1, \dots, d$. Further assume that f_j's are differentiable and N_j's have strictly positive densities.
  - Recall that in defining SCM, the N_j’s are required to be jointly independent, in respect to the independent mechanism principle.

Causal Discovery

Additive Noise Models (ANM).

- Philosophy:



Causal Discovery

Additive Noise Models (ANM).

- All the **non-identifiable cases of ANMs** [Zhang & Hyvarinen, 2009, Thm. 8; Peters et al., 2014, Prop. 23]:
  Consider the bivariate ANM $X \to Y, Y := f(X) + N$, where $N \perp X$ and f is three times differentiable. Assume that $f'(x)(\log p_N)''(y) = 0$ only at finitely many points (x, y).
  If there is a backward ANM $Y \to X$, then one of the followings must hold:

	X	N	f
(i)	Gaussian	Gaussian	linear
(ii)	log-mix-lin-exp	log-mix-lin-exp	linear
(iii), (iv)	log-mix-lin-exp	one-sided asymptotically exponential, or generalized mixture of two exponentials.	strictly monotonic with $\lim_{x \to \infty \text{ or } -\infty} f'(x) = 0.$
(v)	generalized mixture of two exponentials	two-sided asymptotically exponential	strictly monotonic with $\lim_{x \to \infty \text{ or } -\infty} f'(x) = 0.$

Causal Discovery

Common choices of ANMs:

- In multivariate case, even linear Gaussian with equal noise variances is identifiable [Peters and Buhlmann, 2014].
- Linear Non-Gaussian Acyclic Models (LiNGAMs) [Thm. 7.6; Shimizu et al., 2006].
- Nonlinear Gaussian Additive Noise Models [Thm. 7.7; Peters et al., 2014, Cor. 31].

Causal Discovery

Learning causal graph with ANMs.

- Score-based methods.
  - Maximize likelihood using a greedy search technique (for optimization over the graph space).
  - For nonlinear Gaussian ANMs [Buhlmann et al., 2014]:
    - Given a current graph $\mathcal{G}$, regress each var. on its parents and obtain the score $\log p(\mathcal{D}|\mathcal{G})$ $\coloneqq \log p\left(\mathcal{D}\middle|\mathcal{G}, \hat{\theta}\right) = -\sum_{j=1}^{d} \widehat{\mathrm{var}}\left[R_j\right]$, where $R_j \coloneqq X_j - \hat{f}_j\left(\mathrm{pa}_j\right)$ is the residual.
    - When computing the score of a neighboring graph, utilize the decomposition and update only the altered summand.
    - It can be shown to be consistent [Buhlmann et al., 2014].
    - If the noise cannot be assumed to be Gaussian, one can estimate the noise distr. [Nowzohour and Buhlmann, 2016] and obtain an entropy-like score.

Causal Discovery

Learning causal graph with ANMs.

- RESIT [Mooij et al., 2009; Peters et al., 2014]:

Algorithm 1 Regression with subsequent independence test (RESIT)

Input: I.i.d. samples of a p-dimensional distribution on $(X_1, \ldots, X_p)$
$S := \{1, \ldots, p\}, \pi := [\ ]$
PHASE 1: Determine topological order.
repeat
  for $k \in S$ **do**
    Regress X_k on $\{X_i\}_{i \in S \setminus \{k\}}$.
    Measure dependence between residuals and $\{X_i\}_{i \in S \setminus \{k\}}$.
  end for
  Let k^* be the k with the weakest dependence.
  $S := S \setminus \{k^*\}$
  $\mathrm{pa}(k^*) := S$
  $\pi := [k^*, \pi]$ (π will be the topological order, its last component being a sink)
until $\#S = 0$
PHASE 2: Remove superfluous edges.
for $k \in \{2, \ldots, p\}$ **do**
  for $\ell \in \mathrm{pa}(\pi(k))$ **do**
    Regress $X_{\pi(k)}$ on $\{X_i\}_{i \in \mathrm{pa}(\pi(k)) \setminus \{\ell\}}$.
    if residuals are independent of $\{X_i\}_{i \in \{\pi(1), \ldots, \pi(k-1)\}}$ **then**
      $\mathrm{pa}(\pi(k)) := \mathrm{pa}(\pi(k)) \setminus \{\ell\}$
    end if
  end for
end for
Output: $(\mathrm{pa}(1), \ldots, \mathrm{pa}(p))$

Phase 1: Based on the fact that for each node X_i the corresp. N_i is indep. of all non-desc. of X_i. Particularly, for each sink node X_i, $N_i \perp (\mathbf{X} \setminus \{X_i\})$.

Phase 2: visit every node and eliminate incoming edges until the residuals are not indep. anymore.

Causal Discovery

Learning causal graph with ANMs.

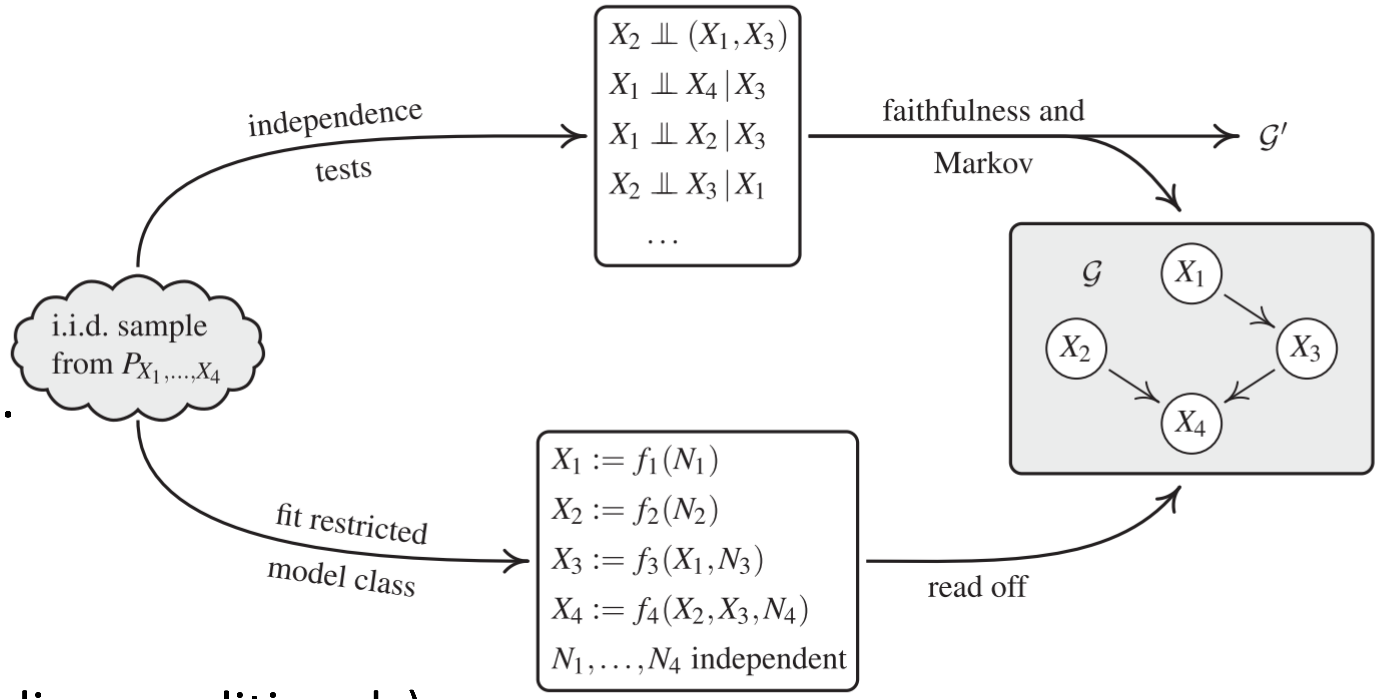
- Independence-Based Score [Peters et al., 2014]:
  - RESIT is asymmetric, thus mistakes will propagate and accumulate over the iterations.
  - Intuition: (i) under $\mathcal{G}_0$, all the residuals are indep. of each other; (ii) causal minimality helps identifying a unique DAG:

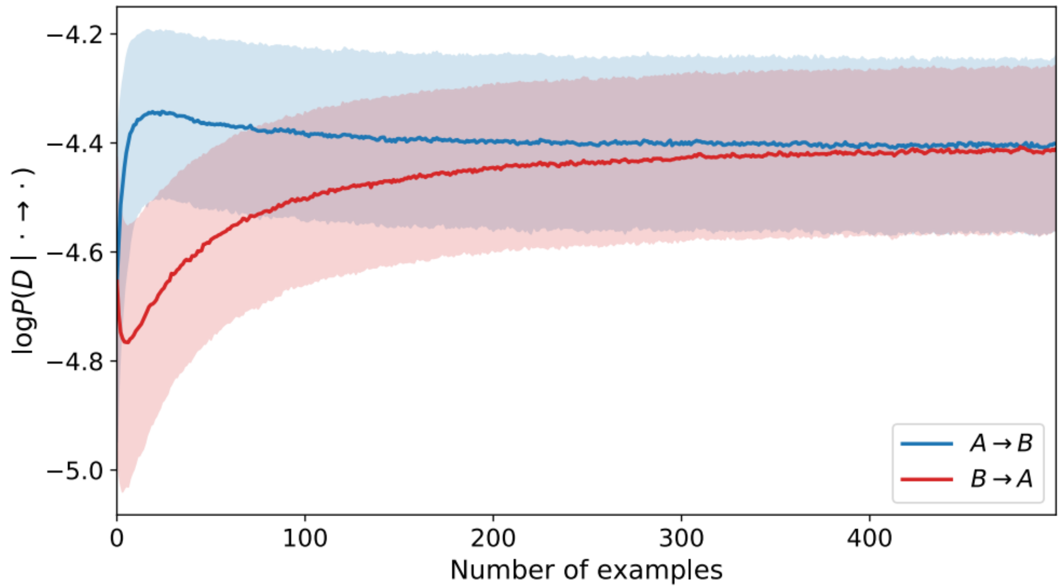
$$\hat{\mathcal{G}} = \underset{\mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{p} \mathrm{DM}(\mathrm{res}_i^{\mathcal{G},\mathrm{RM}}, \mathrm{res}_{-i}^{\mathcal{G},\mathrm{RM}}) + \lambda \,\#(\mathrm{edges})\,. \tag{9}$$

  - $\mathrm{res}_i^{\mathcal{G},\mathrm{RM}}$ are the residuals of node X_i when it is regressed on its parents specified in $\mathcal{G}$ using regression method RM.
  - DM denotes a dependence measure. The second term encourages causal minimality.
- ICA-based methods for LiNGAMs [Shimizu et al., 2006; 2011].

Causal Discovery

- CI-based methods.
  - More faithful to data.
  - Identif. only up to a Markov equiv. class.
  - CI test methods may be tricky.
- ANMs.
  - A relative strong assumption/belief
    (a family is particularly natural for modeling conditionals).
  - Identification within a Markov equiv. class.
- Leveraging intervened datasets.
  - Based on the invariance principle of causality.
  - Identification within a Markov equiv. class.
  - Stronger requirement on data.
  - [Peters et al., 2016; Romeijn & Williamson, 2018; Bengio et al., 2019].





Thanks!