Causal inference cheat sheet

Author: Juvid Aryaman

Last compiled: March 26, 2021

This document is a summary of useful/interesting results in causal inference, mainly from Pearl (2009).

1. Basic probability

- Law of total probability: $P(A) = \sum_i P(A, B_i)$ (a.k.a. marginalizing over B)
- Chain rule of probability: $P(A, B) = P(A|B)P(B)$
- Thus, $P(A) = \sum_i P(A|B_i)P(B_i)$
- Expectation: $E(g(X)) = \sum_x g(x)P(x)$
- Conditional mean: $E(X|Y) = \sum_x xP(x|y)$
- Variance: $\sigma_X^2 = E[(X - E(x))^2]$
- Covariance: $\sigma_{XY} = E[(X - E(X))(Y - E(Y))]$
- Correlation coefficient: $\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$
- Regression coefficient of Y on X: $r_{YX} = \rho_{XY}\sigma_Y/\sigma_X = \sigma_{XY}/(\sigma_X^2)$ (for the equation $Y = r_{YX}X + c + \mathcal{N}(0, \sigma^2)$)
- Conditional independence: $(X \perp\!\!\!\perp Y|Z) \iff P(x|y, z) = P(x|z)$
- Partial correlation $\rho_{XY \cdot Z}$: The correlation between residuals e_X and e_Y resulting from the linear regression of X with Z and Y with Z, respectively.

2. Bayesian networks

Let a **graph** G consist of a set of **vertices** (or **nodes**) V and a set of **edges** E that connect some pair of vertices. Each edge in a graph can be either directed, undirected, or bidirected. Bidirected edges will subsequently be used to denote unobserved common causes, or **confounders**. Let a **path** be a sequence of edges such that each edge starts with the vertex ending in the preceding edge. A path may go either along or against the arrows of a directed graph. Directed graphs may include cycles (e.g. $X \to Y$, $Y \to X$), which represent mutual causation or feedback processes, but not self-loops (e.g. $X \to X$).

The recursive decomposition of the joint distribution into parents which characterises Bayesian networks is

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa_i). \tag{2.1}$$

2.1. d-separation (blocking) in Bayesian networks

Let an arrow $pa_j \to x_j$ denote part of a directed acyclic graph (DAG) in which variables are represented by nodes, and arrows are drawn from each node of the parent set PA_j towards the child node X_j.

Definition 2.1 (d-separation). *A path p is d-separated (or blocked) by a set of nodes Z if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or*
2. *p contains a collider $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z*

A set Z d-separates X from Y if and only if Z blocks **every path** *from a node in X to a node in Y*

Theorem 2.1 (Probabilistic implications of d-separation). *Consequently, if X and Y are d-separated by Z in a DAG G, then $(X \perp\!\!\!\perp Y|Z)$ in every distribution compatible with G. Conversely, if X, Y, and Z are not d-separated by Z in a DAG G then X and Y are dependent conditional on Z in almost all distributions compatible with G (assuming no parameter fine-tuning).*

3. Functional causal models

A functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, ..., n \tag{3.1}$$

where pa_i are the set of variables (parents) that directly determine the value of X_i (i.e. its direct causes) and U_i represents errors (or "disturbances") due to omitted factors. Eq.(3.1) is called a causal model if each equation represents the process by which the *value* (not merely the probability) of variable X_i is selected (see also Definition 11.1).

When some disturbances U_i are judged to be dependent, it is customary to denote such dependencies in a causal graph with double-headed arrows. If the causal diagram is acyclic, then the corresponding model is called *semi-Markovian* and the values of the variables X are uniquely determined by those of the variables U. If the error terms U are jointly independent, the model is called *Markovian*.

Linear structural equation models obey

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, ..., n \tag{3.2}$$

In linear models, pa_i corresponds to variables on the r.h.s. of the above equation where $\alpha_{ik} \neq 0$.

3.1. Counterfactuals in functional causal models: An example

Consider a randomized clinical trial, where patients are/are not treated $X \in \{0, 1\}$. We also observe whether the patients die after treatment $Y\{0, 1\}$. We wish to ask the question: did the patient die *because of* the treatment, *despite* the treatment, or *regardless* of the treatment.

Assume $P(y|x) = 0.5$, and therefore $P(y, x) = 0.25$ for all x and y. We can write two models with the same joint distribution

Model 1 (treatment no effect):

$$x = u_1 \tag{3.3}$$

$$y = u_2 \tag{3.4}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{3.5}$$

Model 2 (treatment has an effect):

$$x = u_1 \tag{3.6}$$

$$y = xu_2 + (1 - x)(1 - u_2) \tag{3.7}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{3.8}$$

Let Q=fraction of deceased subjects from the treatment group who would not have died had they not taken the treatment. In model 1, $Q = 0$ since X has no effect on Y. In model 2, subjects who died ($y = 1$) and were treated ($x = 1$) must correspond to $u_2 = 1$. If $u_2 = 1$ then the only way for $y = 0$ is for $x = 0$. I.e. if you are a patient for whom $u_2 = 1$ then the only way not to die is to not take the treatment, so the treatment caused your death. So $Q = 1$.

Consequence 0: joint probability distributions are insufficient for counterfactual computation

Consequence 1: stochastic causal models are insufficient for counterfactual computation

Consequence 2: functional causal models are sufficient to define and compute counterfactual statements.

3.2. General method to compute counterfactuals

Given evidence $e = \{X_{obs}, Y_{obs}\}$, to compute probability of $Y = y$ under hypothetical condition $X = x$ apply the following steps:

1. Abduction: Update the probability of disturbances $P(u)$ to obtain $P(u|e)$
2. Action: Replace the equations corresponding to variables in the set X by the equations $X = x$
3. Prediction: Use the modified model to compute the probability $Y = y$.

4. Causal Bayesian networks

Theorem 4.1 (Causal Markov condition). *Every Markovian causal model induces a distribution $P(x_1, ..., x_n)$ that satisfies the condition that each variable X_i is independent of all its nondescendants, given its parents PA_i*

Definition 4.1 (Stability). *Stability (or faithfulness), is the assumption that all independencies embedded in a distribution P are entailed by the structure of the causal model D (via Theorem 2.1), and hence remain invariant to any change in the parameters Θ_D.*

The assumption of stability is, in some sense, the inverse of the causal Markov condition. Stability is not always a reasonable assumption to make, especially in evolutionary contexts. Consider the DAG

$$G_A \overset{-}{\to} G_B \overset{+}{\to} P \overset{+}{\leftarrow} G_A \tag{4.1}$$

where G is a gene and P is a protein, and the overset denotes the sign of the direct effect. If these two pathways are usually finely balanced in order to control some critical quantity of P, then it may appear statistically that $G_A \perp\!\!\!\perp P$, but this is an entirely parametric consequence, not a causal one. This is an example of a "back-up" mechanism.

Definition 4.2 (Causal effect). *Given two disjoint sets of variables X and Y, the* **causal effect** *of X on Y, denoted as $P(y|\hat{x})$ or $P(y|do(x))$, is the probability of $Y = y$ by deleting all equations from Eq.*(3.1) *where variables X are on the l.h.s., and substituting $X = x$ in the remaining equations. This corresponds to mutilating the DAG such that all arrows pointing directly to X are removed.* **Amputation is the difference between seeing and doing**.

For an atomic intervention, we get the *truncated factorization* formula

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} \prod\limits_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{4.2}$$

The $j \neq i$ denotes the removal of the term $P(x_i|pa_i)$ from Eq.(2.1) (i.e. amputation). A $do(x_i)$ is a severely limited sub-space of the full joint distribution, since the distribution only has support where the intervention variable x_i is equal to its particular intervention value x'_i, rather than a continuum of values in Eq.(2.1).

Multiplying and dividing by $P(x'_i|pa_i)$ yields

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} P(x_1, ..., x_n|x'_i, pa_i)P(pa_i) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{4.3}$$

Marginalization of the above leads to the following theorem.

Theorem 4.2 (Adjustment for direct causes). *Let PA_i denote the set of direct causes of variable X_i, and let Y be any set of variables disjoint of $\{X_i \cup PA_i\}$. The causal effect of $do(X_i = x'_i)$ on Y is*

$$P(y|\hat{x}'_i) = \sum_{pa_i} P(y|x'_i, pa_i)P(pa_i) \tag{4.4}$$

where $P(y|x'_i, pa_i)$ and $P(pa_i)$ are preintervention probabilities. This is called "adjusting for PA_i".

4.1. Identifiability

Causal quantities are defined relative to a causal model M, not the joint distribution $P_M(v)$ over the set of observed variables V. Non-experimental data provides information about $P_M(v)$ alone, and several graphs can give rise to the same $P_M(v)$. Thus, not all quantities are unambiguously **identifiable** from observational data, **even with infinite samples**. Added assumptions by specifying a particular M can provide enough details to compute quantities of interest without explicating M in full.

Theorem 4.3 (Identifiability). *Given a causal diagram G of any Markovian model in which a subset of variables V are measured, the causal effect $P(y|\hat{x})$ is identifiable whenever $\{X \cup Y \cup PA_X\} \subseteq V$. I.e. when all parents of the cause are measured, the causal effect can be identified.*

5. Inferring causal structure

Definition 5.1 (Pattern). *A pattern is a partially directed DAG in which directed edges are edges common to every member of an equivalence class, and undirected edges represent ambivalence with respect to the direction of the edge.*

5.1. Pattern search algorithms

- IC algorithm is for inferring patterns given observational data when there are no latent variables (Markov)
- IC* algorithm is for inferring causal structure given observational data when there are latent common causes (semi-Markov)
- PC algorithm (Spirtes *et al.*, 2010) assumes that the underlying causal structure of the input data is acyclic, entirely continuous or entirely discrete, and that no two variables are caused by the same latent variable (see Definition 10.2). Note that the appropriate significance threshold is a function of the sample size, and there is no general theory for how the threshold should be dropped with sample size.
- FGS (fast greedy equivalence search), which searches for the whole pattern, and seeks to find a pattern with maximized score. Can allow you to infer the pattern with 10^9 variables or more (Ramsey *et al.*, 2017).
- Lingam (linear non-Gaussian acyclic model for causal discovery) will allow discovery of a **unique** DAG when all relationships are linear but the disturbance terms on all variables are non-Gaussian, with skew (Shimizu *et al.*, 2006).

These algorithms are often point-wise consistent (satisfying assumptions, will converge with probability 1 to an equivalence class containing the true DAG) but not uniformly consistent (can compute error bounds and confidence intervals for finite samples). Other points:

- There are local criteria for potential cause and genuine cause
- Spurious association: X and Y are spuriously associated if they are dependent in some context and there exists a latent common cause, as exemplified in the structure $Z_1 \rightarrow X \rightarrow Y \leftarrow Z_2$

6. Adjusting for confounding bias

When seeking to evaluate the effect of one factor (X) on another (Y), we should ask **whether** we should *adjust* for possible variations in other factors (Z, known as “covariates”, “concomitants” or “confounders”). This becomes apparent in **Simpson’s paradox**: any statistical relationship between two variables may be reversed by including additional factors in the analysis (see Section 10).

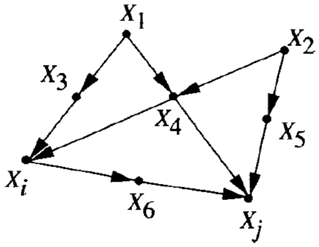


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 1. Example of the back-door criterion

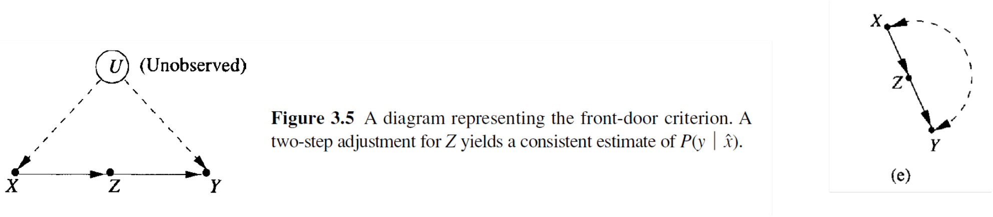


Figure 3.5 A diagram representing the front-door criterion. A two-step adjustment for Z yields a consistent estimate of $P(y \mid \hat{x})$.

Figure 2. (Left) Example of the front-door criterion. The path $X \leftarrow U \rightarrow Y$ denotes an unobserved (latent) unobserved common cause. (Right) This is often represented as a **bi-directed path**.

6.1. The back-door criterion

This criterion demonstrates how confounders that *affect* the treatment variable can be used to facilitate causal inference.

Definition 6.1 (Back-door criterion). *A set of variables Z satisfy the back-door criterion relative to an ordered pair of variables (X_i,X_j) in a DAG 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 G, then Z satisfies the 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 $X_j \in Y$.

Theorem 6.1 (Back-door adjustment). *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*

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z). \tag{6.1}$$

This corresponds to partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X on Y in each homogeneous group, and then averaging the results. Conditioning in this way means that the observation $X = x$ cannot be distinguished from an intervention $do(x)$.

6.2. The front-door criterion

This criterion demonstrates how confounders that are *affected by* the treatment variable can be used to facilitate causal inference.

Definition 6.2 (Front-door). *A set of variables Z satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if:*

1. *Z intercepts all directed paths from X to Y;*
2. *there is no unblocked back-door path from X to Z; and*
3. *all back-door paths from Z to Y are blocked by X.*

Theorem 6.2 (Front-door adjustment). *If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by*

$$P(y|\hat{x}) = \sum_{z} P(z|x) \sum_{x'} P(y|x', z) P(x') \tag{6.2}$$

Conditions (2) and (3) of the front-door definition are overly restrictive: e.g. nested combinations of back-door and front-door conditions are permissible: see Section 7 for a more general set of conditions.

7. Do-calculus

The back-door and front-door criteria do not provide a complete set of rules for when/how causal effects can be computed. Do-calculus sidesteps the need for algebraic manipulation and provides a complete set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, allowing a method of deriving/verifying claims about interventions. The aim is to compute causal effect expressions for $P(y|\hat{x})$ where Y and X are subsets of variables. When $P(y|\hat{x})$ can be reduced to an expression involving observable probabilistic quantities, we say that the causal effect of X on Y is **identifiable**.

7.1. Notation

- $G_{\overline{X}}$ = graph obtained by deleting from G all arrows pointing into nodes in X
- $G_{\underline{X}}$ = graph obtained by deleting from G all arrows pointing out of nodes in X
- $G_{\overline{X}\underline{Z}}$ = graph obtained by deleting from G all arrows pointing into nodes in X and out of nodes in Z
- $P(y|\hat{x}, z) := P(y, z|\hat{x})/P(z|\hat{x})$, meaning the probability of observing $Y = y$ given an *intervention* $X = x$ and an *observation* $Z = z$

7.2. Rules

Rule 7.1 (Insertion/deletion of observations).

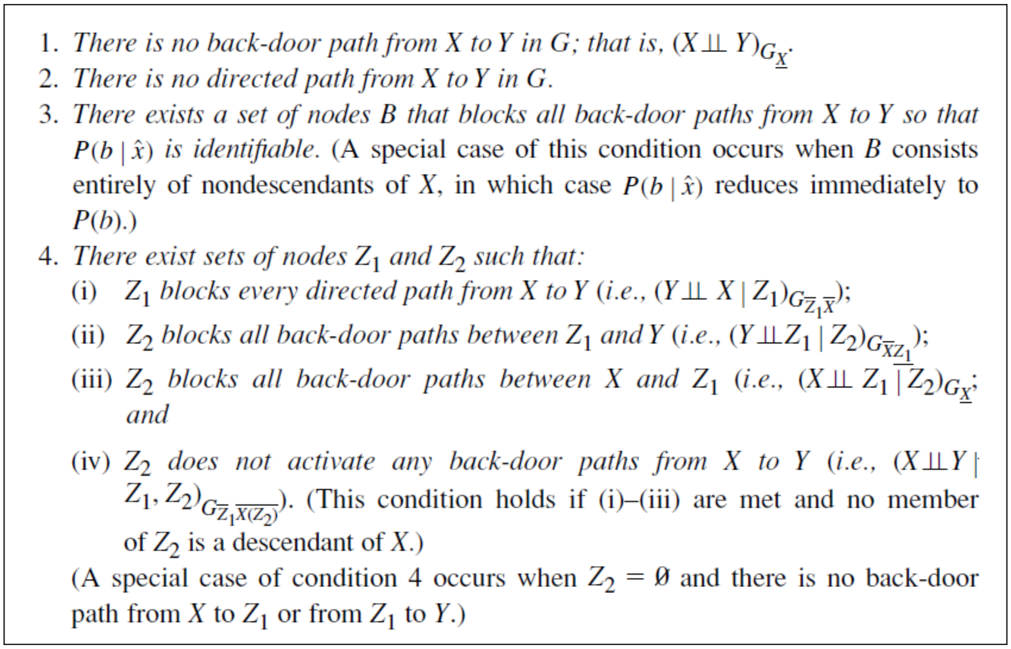
$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{7.1}$$

Rule 7.1 is a reaffirmation of d-separation (Section 2.1) as a valid test for conditional independence in the distribution resulting from $do(X = x)$. The rule follows from the fact that deleting equations from the system ($G_{\overline{X}}$) does not introduce any dependencies among the remaining disturbance terms.

Rule 7.2 (Action/observation exchange).

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{7.2}$$

Rule 7.2 provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.



1. *There is no back-door path from X to Y in G; that is,* $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$.
2. *There is no directed path from X to Y in G.*
3. *There exists a set of nodes B that blocks all back-door paths from X to Y so that* $P(b \mid \hat{x})$ *is identifiable.* (A special case of this condition occurs when B consists entirely of nondescendants of X, in which case $P(b \mid \hat{x})$ reduces immediately to $P(b)$.)
4. *There exist sets of nodes* Z_1 *and* Z_2 *such that:*
   (i) Z_1 *blocks every directed path from X to Y (i.e.,* $(Y \perp\!\!\!\perp X \mid Z_1)_{G_{\overline{Z_1}\,\overline{X}}}$);
   (ii) Z_2 *blocks all back-door paths between* Z_1 *and Y (i.e.,* $(Y \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\overline{X}\underline{Z_1}}}$);
   (iii) Z_2 *blocks all back-door paths between X and* Z_1 *(i.e.,* $(X \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\underline{X}}}$; *and*
   (iv) Z_2 *does not activate any back-door paths from X to Y (i.e.,* $(X \perp\!\!\!\perp Y \mid Z_1, Z_2)_{G_{\overline{Z_1}\,\overline{X(Z_2)}}}$). (This condition holds if (i)–(iii) are met and no member of Z_2 is a descendant of X.)

   (A special case of condition 4 occurs when $Z_2 = \emptyset$ and there is no back-door path from X to Z_1 or from Z_1 to Y.)

Figure 3. Graphical conditions for identification of causal effect (Theorem 4.3.1 Causality). Satisfying at least one renders the causal effect $P(y|\hat{x})$ identifiable, whereas satisfying none implies unidentifiability of the causal effect.

Rule 7.3 (Insertion/deletion of actions).

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}, \overline{Z(W)}}} \tag{7.3}$$

where $Z(W)$ *is the set of* Z*-nodes that are not ancestors of any* W*-node in* $G_{\overline{X}}$.

Rule 7.3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence $G_{\overline{XZ}}$).

Completeness A quantity $Q = P(y|do(x), z)$ is identifiable if and only if it can be reduced to a *do*-free expression using the above 3 rules.

7.3. Identifiability

A causal effect $q = P(y_1, \ldots, y_k|\hat{x}_1, \ldots, \hat{x}_m)$ is identifiable in a model characterised by a graph G if there exists a finite sequence of transformations conforming to one of the three rules in Section 7.2 that reduces q into a standard (i.e. "hat"-free) probability expression involving observed quantities. Figure 3 provides a set of graphical conditions; if any one is satisfied then $P(y|\hat{x})$ is identifiable, and satisfying at least one of the conditions is necessary for $P(y|\hat{x})$ to be identifiable. I.e. $P(y|\hat{x})$ is unidentifiable then no finite sequence of inference rules reduces $P(y|\hat{x})$ to a hat-free expression. Figure 3 can also be used to define an algorithm for deriving a closed-form expression for control queries in terms of observable quantities, see Section 13.2 for an implementation.

Assorted facts on identifiability

- Whilst a causal effect is not identifiable for *every* joint distribution of variables if this condition is broken, it might be for *some* probability densities. For example, an instrumental variable can yield a causal effect identifiable in a linear model in the the presence of a bow pattern (Fig. 3.7A of Causality), but will not be generally identifiable (see Section 3.5 of Causality).

- If $P(y|\hat{x})$ is identifiable, then if a set of nodes Z lies on a directed path from X to Y, then $P(z|\hat{x})$ is also identifiable (lemma 4.3.4).

Definition 7.1 (d-separation equivalence). *For two nodes a and b in a causal diagram M, we write $a\underset{M}{-\!\!-}b$ if a and b are adjacent in model M. We write conditional adjacency $a\underset{M}{-\!\!-}b|c$ to mean once given c, a and b cannot be independent: graphically, a and b are the common parents of c, or some ancestor of c, or a and b are unconditionally adjacent. Two causal models D and E over the set of nodes N, are d-separation equivalent if (Verma and Pearl, 1988):*

$$\underset{a,b\in N}{\forall} a\underset{D}{-\!\!-}b \iff a\underset{E}{-\!\!-}b \tag{7.4}$$

$$\underset{a,b\in N}{\forall} a\underset{D}{-\!\!-}b|c \iff a\underset{E}{-\!\!-}b|c \tag{7.5}$$

Corollary 7.1. *If models D and E are d-separation equivalent (see Definition 7.1), then they are observationally equivalent: i.e. they are unidentifiable from each other.*

Theorem 7.1 (Complete identifiability condition). *A sufficient condition for identifying the causal effect $P(y|do(x))$ is that there exists no bi-directed path (i.e. a path composed entirely of bi-directed arcs, see Fig. 2) between X and any of its children (Tian and Pearl, 2002). Prior to applying this criterion, all nodes which are not ancestors of Y are deleted from the graph (i.e. only consider nodes which are on pathways from X to Y).*

7.4. Confounders, latent common causes, and bi-directed arcs

Bi-directed arcs are, by definition, equivalent to a latent common cause (also known as a confounder) between the two connecting variables (see Fig. 4). This notation is useful because it is the only triad with a single latent variable where the latent common cause renders the causal effect unidentifiable. In Fig. 4B, even if U were available, conditioning upon it would induce collider bias. In Fig. 4C, it is intuitive that U is unnecessary to be known, because all causal effects are mediated by a very large number of unobserved causes, but we do not need complete knowledge of the world to make causal inferences.

8. Actions, plans, and direct effects

Pearl defines two kinds of intervention:

- Act: An intervention which results from a reactive policy, deriving from an agent's beliefs, disposition, and environmental inputs (or the "outside")
- Action: An intervention which results from a deliberative policy, deriving from an agent's free will (or the "inside"; meditative traditions might not draw such a bright line between these two classifications as a description of physical reality, but it is no doubt a useful distinction for reasoning about the future when conscious agents are involved)

8.1. Conditional actions and stochastic policies

In general, interventions may involve complex policies in which X is made to respond according to e.g. a deterministic functional relationship $x = g(z)$, or more generally through a stochastic relationship whereby X is set to x with probability $P^*(x|z)$.

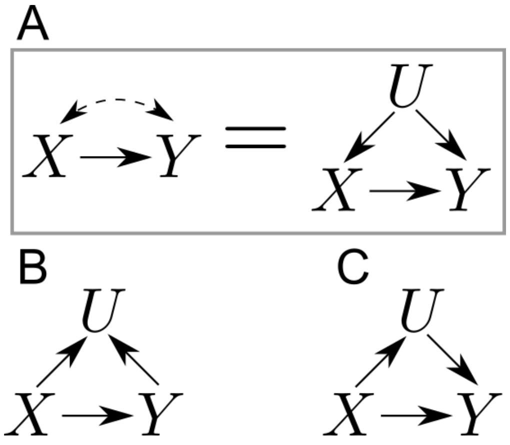


Figure 4. Bi-directed arcs are latent common causes by definition, and are the only acyclic triad with an unobserved variable where the causal effect is unidentifiable. A) A bi-directed arc between variables X and Y is, by definition, equivalent to the graph $U \rightarrow X \rightarrow Y \leftarrow U$ where U is unobserved (latent). This notation is useful because this triad is unique, in the sense that it is the only possible triad involving U where $P(Y|do(x))$ is unidentifiable. B) For this DAG, applying Theorem **??**, $P(Y|do(x)) = P(Y|X)$, and so U is irrelevant for estimation of the causal effect. C) Again, $P(Y|do(x)) = P(Y|X)$ so U is irrelevant. The final permutation of arrows is omitted because it induces a cycle, which is beyond the scope of basic causal inference.

Let $P(y|do(X = g(z)))$ denote the distribution of Y prevailing under the deterministic policy $do(x = g(z))$. Then,

$$\begin{aligned} P(y|do(X = g(z))) &= \sum_z P(y|do(X = g(z)), z)P(z|do(X = g(z))) \\ &= \sum_z P(y|\hat{x}, z)|_{x=g(z)} P(z) \\ &= E_z[P(y|\hat{x}, z)|_{x=g(z)}]. \end{aligned} \tag{8.1}$$

Hence, the evaluation of the outcome of an intervention under a complicated conditional policy $x = g(z)$ amounts to being able to evaluate $P(y|\hat{x}, z)$. The equality $P(z|do(X = g(z))) = P(z)$ stems from the fact that Z **cannot** be a descendant of X: in other words, **one cannot define a coherent policy of action for X based on an (indirect) effect of X because actions change the distributions of their effects!** (Aside: I suppose one might argue about whether an agent has any choice over the form of $g(z)$)

Similarly, let $P(y)|_{P^*(x|z)}$ denote the distribution of Y prevailing under the stochastic policy $P^*(x|z)$ – i.e. given $Z = z$, $do(X = x)$ occurs with probability $P^*(x|z)$. Then,

$$P(y)|_{P^*(x|z)} = \sum_x \sum_z P(y|\hat{x}, z)P^*(x|z)P(z). \tag{8.2}$$

Since $P^*(x|z)$ is specified externally, it is again the case that $P(y|\hat{x}, z)$ is sufficient for the identifiability of any stochastic policy which shapes the distribution of X by the outcome of Z.

8.2. Identification of dynamic plans

A **control problem** consists of a DAG with vertex set V partitioned into four disjoint sets $V = \{X, Z, U, Y\}$ where

- X = the set of control variables (exposures, interventions, treatments, etc.)

- $Z =$ the set of observed variables, often called **covariates**
- $U =$ the set of unobserved (latent) variables, and
- $Y =$ an outcome variable

We are interested in settings where we have gathered data $\mathcal{D} = \{X, Z, Y\}$ for previous agents making actions X. The problem is, given a new instance of the system (e.g. a new patient whom we seek to treat), can we estimate the outcome of $\{do(x_1), ..., do(x_n)\}$ using only the observational data $\mathcal{D}$. See Section 4.4.1 of Causality for a specific motivating example.

Let control variables be ordered $X = X_1, ..., X_n$ such that every X_k is a non-descendant of X_{k+j} $(j > 0)$ and let the outcome Y be a descendant of X_n. A **plan** is an ordered sequence $(\hat{x}_1, ..., \hat{x}_n)$ of value assignments to the control variables. A **conditional plan** is an ordered sequence $(\hat{g}_1(z_1), ..., \hat{g}_n(z_n))$ where $\hat{g}_k(z_k)$ means "set X_k to $\hat{g}_k(z_k)$ whenever $Z_k = z_k$", where the support Z_k of each $g_k(z_k)$ must not contain any variables that descendants of X_k.

Theorem 8.1 (Plan identification: the sequential back-door criterion). *The probability of the* **unconditional** *plan $P(y|\hat{x_1}, ..., \hat{x}_n)$ is identifiable if, for every $1 \leq k \leq n$ there exists a set Z_k of covariates satisfying the following conditions:*

$$Z_k \subseteq N_k \tag{8.3}$$

where N_k is the set of observed nodes that are non-descendants of any element of $\{X_k, X_{k+1}, ..., X_n\}$, and

$$(Y \perp\!\!\!\perp X_k | X_1, ..., X_{k-1}, Z_1, ..., Z_k)_{G_{\underline{X}_k, \overline{X}_{k+1}, ..., \overline{X}_n}} \tag{8.4}$$

When these conditions are satisfied, the effect of the plan is given by

$$P(y|\hat{x}_1, ..., \hat{x}_n) = \sum_{z_1, ..., z_n} P(y|z_1, ..., z_n, x_1, ..., x_n) \times \prod_{k=1}^{n} P(z_k|z_1, ..., z_{k-1}, x_1, ..., x_{k-1}) \tag{8.5}$$

8.3. Direct and indirect effects

We are often concerned with the extent to which a variable affects another directly, rather than the total causal effect mediated through all other intervening variables. For example, in cases of sex discrimination, we may be interested in asking the direct effect of an applicant's sex on the outcome of an applicant's job application. In effect, we are concerned with the causal effect of variable X on Y while all other factors in the analysis are held fixed (*Ceteris paribus*).

Definition 8.1 (Direct effect). *The direct effect of X on Y is given by $P(y|\hat{x}, \hat{s}_{XY})$ where $\hat{s}_{XY}$ is the set of all variables in the model except X and Y*

Corollary 8.1. *The direct effect of X on Y is given by $P(y|\hat{x}, \hat{pa}_{Y \setminus X})$ where $pa_{Y \setminus X}$ is any realization of the parents of Y excluding X.*

It is sometimes meaningful to average the direct effect over all levels of $pa_{Y \setminus X}$. To do this, we define the natural direct effect:

Definition 8.2 (Natural direct effect). *The natural direct effect ($DE_{x,x'}(Y)$) is defined as*

$$DE_{x,x'}(Y) := E[Y(x', Z(x)) - E(Y(x))] \tag{8.6}$$

where $Z = pa_{Y \setminus X}$, and $Y(x', Z(x))$ is the value that Y would attain under the counterfactual scenario of $X = x'$, but Z retaining the values under the setting $X = x$.

The natural direct effect involves probabilities of nested counterfactuals, and cannot generally be written in terms of the $do(x)$ operator. However, if certain assumptions of "no confounding" are deemed valid, the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_{z}[E(Y|do(x', z)) - E(Y|do(x, z))]P(z|do(x)) \tag{8.7}$$

which is simply a weighted average of controlled direct effects.

We can also define the indirect effect which quantifies the influence of X on Y through all paths except for the direct path from $X \to Y$.

Definition 8.3 (Indirect effect). *The natural indirect effect ($IE_{x,x'}(Y)$) is defined as*

$$IE_{x,x'}(Y) = E[Y(x, Z(x')) - E(Y(x))] \tag{8.8}$$

We can define

Definition 8.4 (Total effect). *The total effect of X on Y is given by $P(y|do(x))$, namely, the distribution of Y while X is held constant at x and all other variables are permitted to run their natural course.* **Confusingly**, *we also sometimes denote the total effect $TE_{x,x'}(Y)$ as*

$$TE_{x,x'}(Y) \coloneqq E[Y(x') - E(Y(x))] \tag{8.9}$$

TODO: Write $TE_{x,x'}(Y)$ in terms of $P(y|do(x))$? Is it $E(y|do(x')) - E(y|do(x))$?

Theorem 8.2 (Relationship between total effect, direct effect, and indirect effect). *The total effect of a transition is the difference between the direct effect of that transition and the indirect effect of the reverse transition*

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) - IE_{x',x}(Y) \tag{8.10}$$

9. Causality and structural models

Let's rewrite Eq.(3.1) as

$$x_i = f_i(pa_i, \epsilon_i), \quad i = 1, ..., n. \tag{9.1}$$

In general, for the **partial correlation** $\rho_{XY \cdot Z}$,

$$(X \perp\!\!\!\perp Y|Z) \implies \rho_{XY \cdot Z} = 0 \tag{9.2}$$

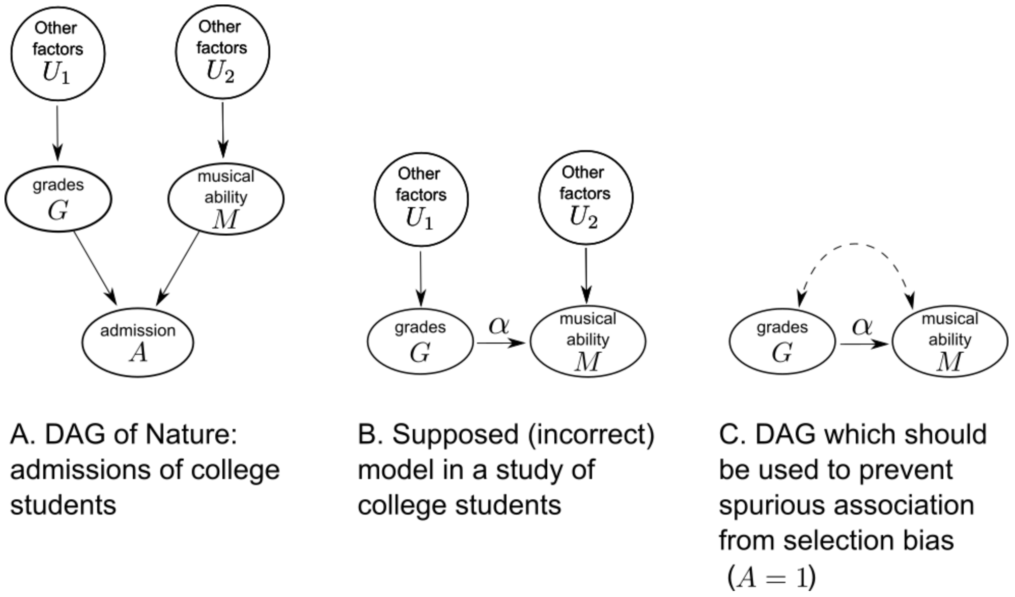
and therefore, in **any** Markovian model with DAG G, the partial correlation $\rho_{XY \cdot Z}$ vanishes whenever the nodes corresponding to the variables in Z d-separate node X from node Y in G, regardless of model parameters. Moreover, no other partial correlation vanishes, for all model parameters. **[Q: Not sure if this is general or only for linear models]**

Theorem 9.1 (Test for correlation between error terms). *For any two non-adjacent variables X and Y, where Y is* **not** *a parent of X, a sufficient test of whether ϵ_X and ϵ_Y are uncorrelated is if the following equality holds:*

$$E[Y|x, do(S_{XY})] = E[Y|do(x), do(S_{XY})] \tag{9.3}$$

where S_{XY} stands for (any setting of) all variables in the model excluding X and Y. If ϵ_X and ϵ_Y are uncorrelated then we are justified in having the absence of a bidirected arc between X and Y. I.e. the omitted factors which directly affect X, ϵ_X, are independent of the omitted factors which directly affect Y, ϵ_Y.

Note that **selection bias** can arise when two uncorrelated factors have a common effect that is omitted from the analysis but influences the selection of samples for the study, see Fig. 5. Hence bidirected arcs should be assumed to exist, by default, between any two nodes in a diagram – since they at worst compromise the identifiability of model parameters. They should be deleted only by well-motivated justifications, such as the unlikely existence of a common cause, and the unlikely existence of selection bias.



A. DAG of Nature: admissions of college students

B. Supposed (incorrect) model in a study of college students

C. DAG which should be used to prevent spurious association from selection bias $(A = 1)$

Figure 5. An example of selection bias in a study. A) Suppose, for a particular college, both grades and musical ability affect admission rate. B) Suppose investigators seek to understand the relationship between grades and musical ability, and measure G and M amongst students of the college in an attempt to estimate α ($\alpha = 0$ in reality). The investigators assume the DAG in (B) and discover a strong negative relationship between G and M. C) However, the investigators would discover that the equality in Eq.(9.3) does not hold, and therefore U_1 and U_2 are correlated (owing to a selection bias from $A = 1$, see Berkson's paradox). Hence the investigators should correct their DAG to include a latent common cause between G and M in order to prevent spurious inferences (see Fig. 4). In doing so, α becomes unidentifiable since all of the causal effect can (and, here, is!) attributable to selection bias.

9.1. Exogeneity

Definition 9.1 (General Exogeneity). *Let X and Y be two sets of variables, and let λ be any quantity which may be computed from a structural model M (structural, statistical, etc.) in a theory T. We say that X is exogenous relative to (Y, λ, T) if λ is identifiable from the conditional distribution of $P(y|x)$, that is, if*

$$P_{M_1}(y|x) = P_{M_2}(y|x) \implies \lambda(M_1) = \lambda(M_2) \tag{9.4}$$

for any two models M_1 and M_2 satisfying theory T.

The following three alternative definitions of exogeneity are also sometimes given, but are not generally equivalent when we cannot assume stability. In order of decreasing strength:

Definition 9.2 (Graphical criterion for exogeneity). *A variable X is exogenous in model M relative to Y if X and Y have no common ancestor in $G(M)$ or, equivalently, if all back-door paths between X and Y are blocked (by colliding arrows). "Common ancestors" should exclude nodes that have no other connection to Y except through X, but should include latent nodes for every pair of dependent errors.*

Definition 9.3 (Error-based criterion for exogeneity). *A variable X is exogenous in model M relative to Y if X is independent of all error terms (U in Defn. 11.1) that have an influence on Y when X is held constant. I.e. A variable X is exogenous relative to $\lambda = P(y|do(x))$ if X is independent of all errors U that influence Y, except those mediated by X.*

Definition 9.4 (Empirical criterion for exogeneity). *A variable X is exogenous relative to Y if and only if*

$$P(Y_x = y) = P(y|x) \tag{9.5}$$

(see Section 11.1 for definitions of counterfactuals), or equivalently

$$P(Y = y|do(x)) = P(y|x) \tag{9.6}$$

where

$$\text{graphical criterion} \implies \text{error-based criterion} \implies \text{counterfactual criterion.} \tag{9.7}$$

However, if we assume stability, then the three definitions coincide.

9.2. Instrumental variables

Definition 9.5 (Instrumental variable). *For the parameter $\lambda = P(w|do(z))$ where W and Z are two sets of variables in Y, then if X is exogenous relative to (Y, λ, T) then*

$$P_{M_1}(z, w|x) = P_{M_2}(z, w|x) \implies P_{M_1}(w|do(z)) = P_{M_2}(w|do(z)). \tag{9.8}$$

for any two models M_1 and M_2 satisfying theory T. Under these conditions, we call X an instrumental variable for the causal effect $P(w|do(z))$

Corollary 9.1. *If X is an instrumental variable for $P(w|do(z))$, then $P(w|do(z))$ is identifiable from $P(z, w|x)$. I.e. X renders a causal effect $P(w|do(z))$ identifiable which X itself does not directly participate in.*

As in the case of exogeneity, we can provide graphical and counterfactual criteria for instrumental variables

Definition 9.6 (Criteria for instrumental variables). *A variable Z is an instrument relative to the total effect of X on Y if there exists a set of measurements $S = s$ (i.e. joint observations for X, Y, and Z?* **[TODO: Confirm?]***), unaffected by X, such that either of the following criteria holds*

1. **Counterfactual** *criterion*
   (a) $Z \perp\!\!\!\perp Y_x | S = s$
   (b) $Z \not\perp\!\!\!\perp X | S = s$
2. **Graphical** *criterion*
   (a) $(Z \perp\!\!\!\perp Y | S)_{G_{\bar{X}}}$
   (b) $(Z \not\perp\!\!\!\perp X | S)_G$

9.3. Linear structural equation models

Linear structural equation models (SEMs) obey

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + \epsilon_i, \quad i = 1, ..., n \tag{9.9}$$

If, in Eq.(9.9)

$$\epsilon \sim \mathcal{N}(\mu, \Sigma) \tag{9.10}$$

then X_i will also be multivariate normal, and the SEM will be entirely determined by the set of correlation coefficients ρ_{ij}. For a linear SEM

$$\rho_{XY \cdot Z} = 0 \iff (X \perp\!\!\!\perp Y | Z). \tag{9.11}$$

9.3.1. Interpretation

Definition 9.7 (Structural equations). *An equation*

$$y = \beta x + \epsilon \tag{9.12}$$

is said to be structural if it is to be interpreted as follows: In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, the value of Y is given by $\beta x + \epsilon$, where ϵ is not a function of the settings x and z. **[I think: ϵ can be arbitrarily distributed, potentially correlated with X, but assumed to have $E[\epsilon] = 0$.]**

The equality sign in structural equations has a different behaviour to algebraic equality signs. In the context of observations, the equality sign in Eq.(9.12) behaves symmetrically between X and Y: e.g. observing $Y = 0$ implies $\beta x = -\epsilon$. In contrast, in the context of interventions, the equality sign in Eq.(9.12) behaves asymmetrically between X and Y: e.g. setting $Y = 0$ tells us nothing about the relationship between x and ϵ.

Furthermore, the strongest empirical claim made by Eq.(9.12) is that

$$P(y|do(x), do(z)) = P(y|do(x)) \tag{9.13}$$

i.e. the statistics of Y remain invariant to the manipulation of Z under the condition of $do(x)$. In contrast, regression equations make no empirical claims whatsoever.

The operational definition of the structural parameter β in Eq.(9.12) is

$$\beta = \frac{\partial}{\partial x} E[Y|do(x)] \tag{9.14}$$

(since $E[Y|do(x)] = \beta x$). In words, β is the rate of change in the expectation of Y in an experiment where X is held at x by external control. This is true regardless of the correlation between X and ϵ in non-experimental studies (e.g. via another equation $x = \alpha y + \delta$).

As a consequence of the above, the operational definition of the error term ϵ is

$$\epsilon = y - E[Y|do(x)] \tag{9.15}$$

(again since $E[Y|do(x)] = \beta x$).

9.3.2. Estimation

Define the conditional variance $\sigma^2_{X|z}$, conditional covariance $\sigma^2_{XY|z}$, and the conditional covariance $\rho_{XY|z}$. For multivariate normal, $\sigma^2_{X|z}$, $\sigma^2_{XY|z}$, and $\rho_{XY|z}$ are all independent of the value of z. For the MVN, the **partial** variance $\sigma^2_{X \cdot Z}$, covariance $\sigma_{XY \cdot Z}$, and correlation $\rho_{XY \cdot Z}$ all coincide with the conditional variance, covariance, and correlation respectively (although this is not generally the case).

A **partial regression coefficient**, $r_{YX \cdot Z}$ is given by

$$r_{YX \cdot Z} = \rho_{YX \cdot Z} \frac{\sigma_{Y \cdot Z}}{\sigma_{X \cdot Z}} \tag{9.16}$$

and is equal to the coefficient of X in the linear regression of Y on X and Z. So, the coefficient of x in the regression equation

$$y = \alpha x + b_1 z_1 + ... + b_k z_k \tag{9.17}$$

is

$$\alpha = r_{YX \cdot Z_1 Z_2 ... Z_k} \tag{9.18}$$

Theorem 9.2 (d-Separation in General Linear Models). *For any linear model Eq.*(9.9)*, which may include cycles and bidirected arcs (i.e. dependent ϵ between different variables), $\rho_{XY \cdot Z} = 0$ if Z d-separates X from Y, where bidirected arcs between i and j are interpreted as a latent common parent $i \leftarrow L \rightarrow j$.*

Theorem 9.2 provides a method for finding models in the context of linear SEMs: by searching over all $\rho_{XY \cdot Z}$, we can construct a DAG. Not all partial correlations need to be searched.

Definition 9.8 (Basis). *Let S be a set of partial correlations. A basis B for S is a set of zero partial correlations where (i) B implies the zero of every element of S and (ii) no proper subset of B sustains such an implication.*

An obvious choice of basis for a DAG D is

$$B = \{\rho_{ij \cdot pa_i} | i > j\} \tag{9.19}$$

where i ranges over all nodes in D and j ranges over all predecessors of i in any order that agrees with the arrows of D. More economical choice of basis exist, such as a Graphical Basis.

Theorem 9.3 (Markov linear-normal equivalence). *Two Markovian linear-normal models are observationally indistinguishable if every covariance matrix generated by one model can be parametrically generated by the other (covariance equivalent). Two such models are covariance equivalent if and only if their corresponding graphs have the same sets of zero partial correlations. Moreover, two such models are covariance equivalent if and only if they have the same edges and the same sets of v-structures. (I.e. arrows can be reversed as long as they do not alter v-structures, see Corollary 7.1).*

9.3.3. Parameter identifiability

Consider an edge $X \rightarrow Y$ in graph G, and let α be the path coefficient associated with that edge (i.e. the strength of the direct causal effect of X on Y). The regression coefficient in a linear model can, in general, be decomposed into

$$r_{YX} = \alpha + I_{YX} \tag{9.20}$$

where I_{YX} is independent of α, since I_{YX} is composed of other indirect paths connecting X and Y. If we remove the edge $X \rightarrow Y$ and observe that the resulting subgraph entails zero correlation between X and Y then $I_{XY} = 0$ and $r_{YX} = \alpha$, and hence α is identified. This idea is extended in the following theorem

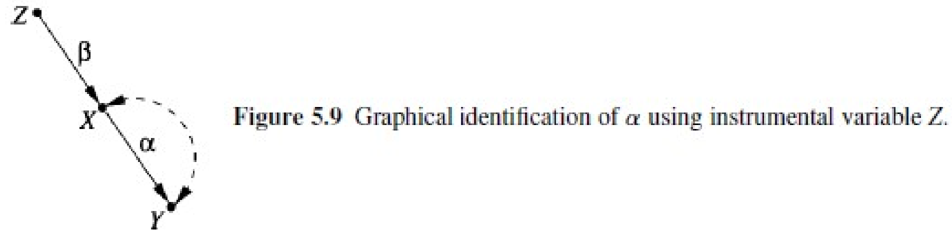


Figure 5.9 Graphical identification of α using instrumental variable Z.

Figure 6. Example of inference of a direct effect by evaluating a broader causal effect, and extracting the effect of interest using an instrumental variable.

Theorem 9.4 (Single-door Criterion for Direct Effects). *Let G be any path diagram in which α is the path coefficient associated with link $X \to Y$, and let G_α denote the diagram that results when $X \to Y$ is deleted from G. The coefficient α is identifiable if there exists a set of variables Z such that (i) Z contains no descendant of Y and (ii) Z d-separates X from Y in G_α. If Z satisfies these two conditions then, in a linear SEM, $\alpha = r_{YX\cdot Z}$. Conversely, if Z does not satisfy these conditions, then $r_{YX\cdot Z}$ is not a consistent estimand of α.*

Theorem 9.5 (Back-door Criterion for Total Effects). *For any two variables in a causal diagram G, the total effect of X on Y is identifiable if there exists a set of measurements Z such that*

1. *no member of Z is a descendant of X; and*
2. *Z d-separates X from Y in the subgraph $G_{\underline{X}}$ formed by deleting all G arrows emanating from X*

If the two conditions are satisfied, then the total effect of X on Y in a linear SEM is given by $r_{YX\cdot Z}$.

Theorems 9.4 and 9.5 are special cases of a more general scheme. In order to identify any **partial effect**, as defined by a select bundle of causal paths from X to Y, we must find a set Z of measured variables that block all non-selected paths between X and Y. For linear models, the partial effect is equal to the regression coefficient $r_{YX\cdot Z}$.

Some direct effects require evaluation of a broader causal effect first, in order to extract the direct effect of interest (see Fig. 6). The parameter α cannot be directly estimated with Theorem 9.4 because of the confounder, or its constituents (since it has none). Instead, we may apply Theorem 9.5 twice like so:

$$P(Y|\hat{z}) = r_{YZ} = \alpha\beta \tag{9.21}$$

$$P(X|\hat{x}) = r_{YX} = \beta \tag{9.22}$$

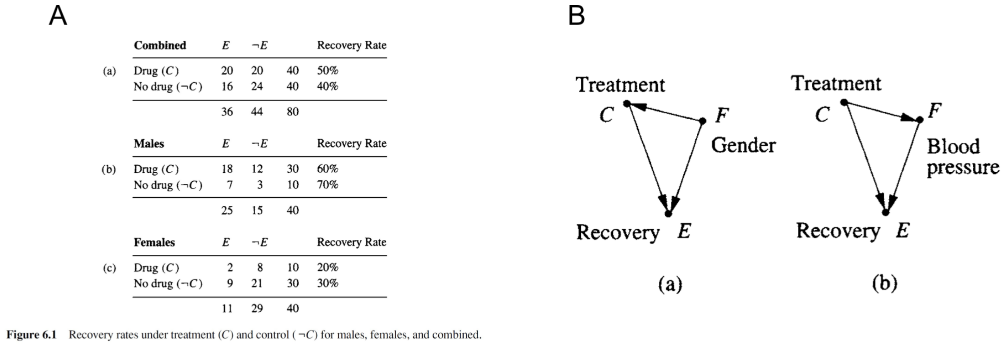
$$\alpha = E(Y|\hat{x}) = \frac{r_{YZ}}{r_{YX}} = \frac{P(Y|\hat{z})}{P(X|\hat{z})} \tag{9.23}$$

10. Simpson's paradox, confounding, and collapsibility

10.1. Simpson's paradox

Simpson's paradox is a reversal effect observed in sub-populations.

Definition 10.1 (Simpson's paradox). *The phenomenon whereby an event C increases the probability of E in a given population p and, at the same time, decreases the probability of E in every sub-population of p. In other words, if F and $\neg F$ are two complementary properties describing two sub-populations,*



	Combined	E	$\neg E$		Recovery Rate
(a)	Drug (C)	20	20	40	50%
	No drug ($\neg C$)	16	24	40	40%
		36	44	80	

	Males	E	$\neg E$		Recovery Rate
(b)	Drug (C)	18	12	30	60%
	No drug ($\neg C$)	7	3	10	70%
		25	15	40	

	Females	E	$\neg E$		Recovery Rate
(c)	Drug (C)	2	8	10	20%
	No drug ($\neg C$)	9	21	30	30%
		11	29	40	

Figure 6.1 Recovery rates under treatment (C) and control ($\neg C$) for males, females, and combined.

Figure 7. Example of Simpson's paradox. A. The recovery rate in the pooled population is greater with the drug, but for both the male and female populations the recovery rate is greater without the drug. B. Two causal models which are observationally equivalent capable of generating the data in (A). In (A), $P(E|do(C)) = \sum_F P(E|C,F)P(F)$ (Theorem 6.1) which averages over sub-populations, whereas in (B), $P(E|do(C)) = P(E|C)$ (Theorem **??**) which uses the pooled data.

then it is possible to encounter the equalities:

$$\begin{aligned} P(E|C) &> P(E|\neg C) \\ P(E|C,F) &< P(E|\neg C,F) \\ P(E|C,\neg F) &< P(E|\neg C,\neg F). \end{aligned} \tag{10.1}$$

Note: Simpson's paradox is not strictly a paradox because it does not involve any contradiction.

An example dataset displaying Simpson's paradox is in Fig. 7 for a drug treatment (C), recovery (E), for sub-populations of gender (F). The question, therefore, is: "what is the total effect of the drug on recovery?". The solution to this question depends on the causal assumptions we bring to bear on the problem: two different causal models which are capable of generating the data yield different answers for the quantity $P(E|do(C))$, one of which averages over the subpopulations, and the other using pooled data.

Theorem 10.1 (Sure-thing Principle). *An action C that increases the probability of an event E in each sub-population F must also increase the probability of E in the population as a whole, provided that the action does not change the distribution of the sub-populations. In other words, for dichotomous sub-populations F, if*

$$\begin{aligned} P(E|do(C),F) &< P(E|do(\neg C),F) \\ P(E|do(C),\neg F) &< P(E|do(\neg C),\neg F) \end{aligned} \tag{10.2}$$

then

$$P(E|do(C)) < P(E|do(\neg C)) \tag{10.3}$$

Theorem 10.1 follows from do-calculus, and gives the intuitive result that if the drug in Fig. 7A harms both men and women then one should not administer the drug if one does not know the patient's gender, despite the fact that the observational conditional densities obey $P(E|C) > P(E|\neg C)$. The "paradox" in the example arises because the males, who recover (regardless of the drug) more often than the females, are also more likely than the females to use the drug.

10.2. Confounding

Definition 10.2 (No confounding). *Variables X and Y are not confounded in a causal model M if and only if*

$$P(y|do(x)) = P(y|x), \text{ or } P(x|do(y)) = P(x|y) \tag{10.4}$$

for all x and y in their respective domains. If this condition holds, we say $P(x|y)$ is unbiased.

Note that no confounding is a special case of exogeneity under a theory T of the set of all models with a given DAG (see Definition 9.1).

Theorem 10.2 (Common-cause principle). *Let A_D be the set of assumptions embedded in an acyclic causal diagram D. Variables X and Y are stably unconfounded given A_D if and only if X and Y have no common ancestor in D.*

The following theorem is an operational test for stable no-confounding (i.e. no confounding for every parametrization of the causal diagram):

Theorem 10.3 (Criterion for stable no-confounding). *Let A_Z denote the assumptions that (i) the data are generated by some (unspecified) acyclic model M and (ii) Z is a variable in M that is unaffected by X, but does contain a directed path from Z to Y, and so Z may possibly affect Y. If both of the following conditions are* **violated***:*

1. *$P(x|z) = P(x)$ (i.e. no association between Z and X)*
2. *$P(y|z, x) = P(y|x)$ (i.e. Z is not associated with Y conditional on X)*

then (X, Y) are **not** *stably unconfounded given A_Z. I.e. finding* **any** *variable Z that satisfies A_Z and violates both of the conditions above permits us to disqualify (X, Y) as stably unconfounded.*

10.3. Collapsibility and confounding

Definition 10.3. *Let $g[P(x, y)]$ be any functional that measures an association between X and Y in the joint distribution $P(x, y)$. We say that g is collapsible on a variable Z if*

$$E_z g[P(x, y|z)] = g[P(x, y)] \tag{10.5}$$

If g stands for any linear functional of $P(y|x)$, such as the risk difference $P(y|x_1) - P(y|x_2)$, then collapsibility holds whenever $Z \perp\!\!\!\perp X$ or $Z \perp\!\!\!\perp Y|X$. Hence, any violation of collapsibility implies violation of the two statistical conditions in Theorem 10.3. So whilst collapsibility is linked to confounding, in the absence of the assumptions A_z, collapsibility is neither necessary nor sufficient.

11. Structure-based counterfactuals

11.1. Definitions

Let's begin by writing Eq.(3.1) more formally:

Definition 11.1 (Structural Causal Model). *A causal model is a triple*

$$M = \langle U, V, F \rangle$$

where

1. *U is a set of background (exogenous, or error) variables, that are determined by factors outside the model;*
2. *V is a set $\{V_1, V_2, ..., V_n\}$ of (endogenous) variables, that are determined by variables in the model – that is, variables in $U \cup V$; and*
3. *F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup PA_i$, where $U_i \subseteq U$ and $PA_i \subseteq V \setminus V_i$ and the entire set F forms a mapping from U to V. In other words, each f_i in*
$$v_i = f_i(pa_i, u_i),\ i = 1, ..., n \tag{11.1}$$
assigns a value to V_i that depends on (the values of) a select set of variables in $V \cup U$, and the entire set F has a unique solution $V(u)$.

Uniqueness is ensured in recursive systems (acyclic, see Defn. 11.10), but multiple solutions are allowed in non-recursive systems. Every causal model M can be associated with a graph $G(M)$ called a **causal diagram** where nodes are variables and directed edges point from members of PA_i and U_i towards V_i. $G(M)$ identifies the endogenous and background variables that have direct influence on each V_i, but not the functional form f_i. The convention of confining the parent set PA_i to variables in V stems from the fact that the background variables are often unobservable.

Definition 11.2 (Submodel). *Let M be a causal model, X a set of variables in V, and x a particular realization of X. A submodel M_x of M is the causal model*

$$M_x = \langle U, V, F_x \rangle,$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{11.2}$$

I.e. F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Definition 11.3 (Effect of action). *Let M be a causal model, X a set of variables in V, and x a particular realization of X. The effect of action $do(X = x)$ on M is given by the submodel M_x.*

Definition 11.4 (Potential response). *Let X and Y be two subsets in V. The potential response of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x. That is, $Y_x(u) = Y_{M_x}(u)$. If Y is a set of variables $Y = (Y_1, Y_2, ...)$ then $Y_x(u) = (Y_{1_x}, Y_{2_x}, ...)$.*

Definition 11.5 (Counterfactual). *Let X and Y be two subsets of variables in V. The counterfactual sentence "Y would be y in situation u, had X been x" is interpreted as the equality $Y_x(u) = y$, with $Y_x(u)$ being the potential response of Y to $X = x$.*

Definition 11.6 (Probabilistic causal model). *A probabilistic causal model is a pair*

$$\langle M, P(u) \rangle$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of the background variables U.

The function $P(u)$, together with f_i, defines a probability distribution over endogenous variables

$$P(y) \coloneqq P(Y = y) = \sum_{\{u|Y(u)=y\}} P(u) \tag{11.3}$$

and counterfactual probabilities are defined in a similar model through the function $Y_x(u)$ induced by submodel M_x:

$$P(Y_x = y) = P(Y = y) = \sum_{\{u|Y_x(u)=y\}} P(u). \tag{11.4}$$

Similarly, one can define joint densities over counterfactuals and observations over (not necessarily disjoint) sets of variables

$$P(Y_x = y, X = x') = \sum_{\{u|Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{11.5}$$

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u|Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u) \tag{11.6}$$

We may interpret statements like "the probability that $X = x$ was the cause of event $Y = y$" using the counterfactual statement "the probability that Y would not be equal to y had it not been for $X = x$, given that $X = x$ and $Y = y$ have in fact occurred". We can write this counterfactual as a conditional on observations

$$P(Y_{x'} = y'|X = x, Y = y) = \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \tag{11.7}$$

$$= \sum_{u} P(Y_{x'}(u) = y')P(u|x, y). \tag{11.8}$$

In other words, we use the observations to update $P(u)$ to $P(u|x, y)$, and then use this update to compute the expectation of the equality $Y_{x'} = u$ [**TODO: Clarify after Ch 9**]. This substantiates the three-step procedure discussed in Section /3.1, summarized in the following theorem:

Theorem 11.1 (Computation of counterfactual statements). *Given model $\langle M, P(u)\rangle$, the conditional probability $P(B_A|e)$ of a counterfactual statement "If it were A then B", given observational evidence e, can be evaluated using the following three steps:*

1. **Abduction** – *Update $P(u)$ by the evidence e to obtain $P(u|e)$. (Abduction means reasoning from evidence to explanation)*
2. **Action** – *Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain submodel M_A.*
3. **Prediction** – *Use the modified model $\langle M_A, P(u|e)\rangle$ to compute the probability of B, the consequence of the counterfactual.*

Definition 11.7 (Connection between causal effect and counterfactual).

$$P(Y|do(x)) \coloneqq P(Y_x = y) \tag{11.9}$$

Definition 11.8 (Worlds and theories). *A causal world w is a pair $\langle M, u\rangle$, where M is a causal model and u is a particular realization of the background variables $U = u$ (i.e. $P(u) = 1$). A causal theory is a set of causal worlds.*

The crucial difference between counterfactual statements and a statement about actions, is that **in counterfactual statements, the facts (i.e. the evidence) can potentially be affected by the antecedents**. For example, in the two-man firing squad example, "If the prisoner is dead [evidence], then the prisoner would be dead even if rifleman A had not shot [antecedent]". In this sentence, the fact that the prisoner is dead could be affected by the fact that A did not shoot. This is not the case for statements about actions. For example, "If the captain gave no signal [evidence] and rifleman A decides to shoot [antecedent], then the prisoner will die and B will not shoot". Here, the evidence that the captain gives no signal is not affected by the antecedent that A shoots. In natural language, counterfactual utterances tend to presume knowledge of facts that are affected by the antecedent. For example "O would be different were it not for A" implies knowledge of the actual value of O, and that O is susceptible to A. In general, some knowledge of the functional mechanisms of $f_i(pa_i, u_i)$ are necessary to evaluate such statements.

In a more general sense, causal effects tend to be concepts based on *general* causes (e.g. "Drinking hemlock causes death"), whereas counterfactuals tend to be concepts based on *singular* causes (e.g. "Socrates' drinking hemlock caused his death"). The latter (counterfactuals) tends to require access to more detailed, mechanistic, specifications and higher computational resources than the former (generic causes). Formally, the difference hinges on whether we need to condition our beliefs on the cause/effect events that actually occurred.

11.2. Twin network representation

A difficulty in Theorem 11.1 is the need to compute, store, and use the posterior distribution $P(u|e)$. This difficulty can be avoided using the twin network representation for counterfactual computation. We

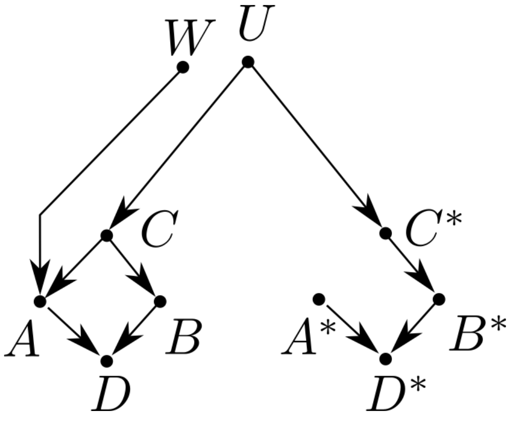


Figure 8. Twin network representation of (semi-deterministic) firing squad problem. A court U issues an order to captain C to execute a prisoner, with some probability. Upon receiving an order, the captain instructs both riflemen A and B to fire. Rifleman A is of a nervous disposition (W), causing him to fire randomly (and independently of the order), with some probability. The prisoner dies if either A or B opens fire. Consider the counterfactual query: "Given that the prisoner is dead, what is the probability that the prisoner would be alive if A did not shoot?", i.e. $P(\neg D_{\neg A}|D)$. This can be answered without reference to $P(u, w|D)$, only the ordinary conditional probability $P(\neg D^{*}|D)$, using the twin network representation.

replicate the endogenous variables and label them distinctly, share the exogenous variables, and amputate edges of the twin network to represent the antecedent of the counterfactual. Given a counterfactual query $P(Y_x = y|z)$ for arbitrary X, Y, and Z, it is sufficient to compute $P(y^{*}|z)$ which can be performed by standard evidence propagation techniques, rather than computing the full posterior distribution $P(u|e)$ – see Fig. 8. The twin representation is also useful for testing independencies between counterfactual quantities by testing d-separation (Defn. 2.1).

11.3. Simon's causal ordering

In general, a physical law need not be specified as a system of structural equations (Eq.11.1), but as a set of functional constraints

$$G_k(x_1, ..., x_l; u_1, ..., u_m) = 0 \tag{11.10}$$

without identifying a 'dependent' variable (positivism: where all descriptions of Nature are defined through 'functiona relations' and 'interdependence' amongst variables). This seems to be at odds with the notion of causation.

Simon (1977) devised a procedure for deciding whether a collection of such functions G_k dictates a unique way of selecting an endogenous dependent variable for each mechanism G_k. The procedure involves finding an order of variables $(X_1, ..., X_n)$ such that we can solve for each X_i without solving for any of X_i's successor. If such an ordering exists, it dictates the direction of causal attribution. The ordering is unique if one can find a unique one-to-one correspondence between equations G and variables X. If the matching is unique, then the choice of dependent variable is unique and defines a DAG (Nayak, 1994). This means that the problem of causal induction can be reduced to the more familiar problem of scientific induction.

11.4. Axioms of structural counterfactuals

We present three properties of counterfactuals – composition, effectiveness, and reversibility – that hold in all causal models.

Property 11.1 (Composition). *For any three sets of endogenous variables X, Y, and W in a causal model, we have*

$$W_x(u) = w \implies Y_{xw}(u) = Y_x(u) \tag{11.11}$$

Composition states that, if we force a variable (W) to a value w that it would have had without our intervention, then the intervention will have no effect on other variables (Y) in the system. That invariance holds for all fixed conditions $do(X = x)$.

Definition 11.9 (Null action).

$$Y_{\emptyset}(u) := Y(u)$$

Corollary 11.1 (Consistency). *For any set of variables Y and X in a causal model,*

$$X(u) = x \implies Y(u) = Y_x(u) \tag{11.12}$$

In other words, if we force variable X to take the value it would have naturally had, then a counterfactual is equivalent to the observational value.

Property 11.2 (Effectiveness). *For all sets of variables X and W,*

$$X_{xw}(u) = x. \tag{11.13}$$

Effectiveness states that if we force a variable X to have the value x, then X will indeed take on the value x.

Property 11.3 (Reversibility). *For any two variables Y and W, and any set of variables X*

$$(Y_{xw}(u) = y) \;\&\; (W_{xy}(u) = w) \implies Y_x(u) = y \tag{11.14}$$

Reversibility precludes multiple solutions due to feedback loops. If setting W to a value w results in a value y for Y, and if setting Y to the value y results in W achieving the value w, then W and Y will naturally obtain the values w and y, without any external setting. This holds for all fixed conditions $do(X = x)$. Reversibility, as an axiom, is only required for non-recursive systems (see Defn. 11.10 below): i.e. directed acyclic graphs only need composition and effectiveness to certify all truth statements.

Properties 11.1, 11.2, and 11.3 hold in all causal models – they are necessary and sufficient axioms for all causal statements. In other words, if we can reduce a counterfactual quantity Q into an expression that involves only ordinary probabilities using these three axioms, then Q is identifiable. If Q cannot be reduced in terms of ordinary probabilities using these three axioms, then Q is unidentifiable. The axioms are as powerful as can be.

Definition 11.10 (Recursiveness). *Let X and Y be singleton variables in a model, and let $X \to Y$ stand for the inequality $Y_{xw}(u) \neq Y_w(u)$ for some values of x, w, and u. A model M is recursive if, for any sequence $X_1, X_2, ..., X_k$, we have*

$$X_1 \to X_2,\ X_2 \to X_3,\ ...,\ X_{k-1} \to X_k \implies X_k \nrightarrow X_1 \tag{11.15}$$

Properties 11.1, 11.2, and Defn. 11.10 are complete: i.e. in a recursive system, only composition and effectiveness are needed to derive all causal statements. Property 11.3 is only required in non-recursive systems.

11.4.1. Rules of inference with counterfactuals (potential outcomes framework)

The following two rules can be used to embody recursiveness

Rule 11.1 (Exclusion restrictions). *For every variable Y having parents PA_Y and for every set of variables $Z \subset V$ disjoint of PA_Y, we have*

$$Y_{pa_Y}(u) = Y_{pa_Y,z}(u) \tag{11.16}$$

Rule 11.1 reflects the insensitivity of Y to any manipulation in V, once its direct causes PA_Y are held constant.

Rule 11.2 (Independence restrictions). *If $Z_1, ..., Z_k$ is any set of nodes in V not connected to Y via paths containing only U variables, we have*

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1_{pa_{Z_1}}}, ..., Z_{k_{pa_{Z_k}}}\} \tag{11.17}$$

where $Z_{i_{pa_{Z_i}}}$ is the potential response of the variable Z_i to setting the parents of Z_i to the value pa_{Z_i}. Equivalently, Eq. 11.17 holds if the corresponding U terms $(U_{Z_1}, ..., U_{Z_k})$ are jointly independent of U_Y.

Rule 11.2 interprets independencies among U variables as independencies between the counterfactuals of the corresponding V variables. This is because, since $Y = f_y(pa_Y, u_Y)$, holding PA_Y fixed means that the residual variations of Y are entirely governed by variations in U_Y.

Using Rules 11.1, 11.2, and Properties 11.1 and 11.2, can be used to compute any causal query in a DAG.

11.4.2. Causal relevance

Causal relevance is concerned with questions of the form: "Given that Z is fixed, wold changing X alter Y?". Causal relevance can be useful when exact causal models do not exist, but constraints about the lack of influence of certain variables on others are known.

Definition 11.11 (Causal irrelevance). *A variable X is causally irrelevant to Y, given Z (written $X \not\twoheadrightarrow Y|Z$) if, for every set W disjoint of $X \cup Y \cup Z$, we have*

$$\forall(u, z, x, x', w), \quad Y_{xzw}(u) = Y_{x'zw}(u) \tag{11.18}$$

where x and x' are two distinct values of X.

Theorem 11.2. *For any causal model, the following sentences must hold.*

Weak Right Decomposition:

$$(X \not\twoheadrightarrow YW|Z) \;\&\; (X \not\twoheadrightarrow Y|ZW) \implies (X \not\twoheadrightarrow Y|Z) \tag{11.19}$$

Left Decomposition:

$$(XW \not\twoheadrightarrow Y|Z) \implies (X \not\twoheadrightarrow Y|Z) \;\&\; (W \not\twoheadrightarrow Y|Z) \tag{11.20}$$

Strong Union:

$$(X \not\twoheadrightarrow Y|Z) \implies (X \not\twoheadrightarrow Y|ZW)\ \forall W \tag{11.21}$$

Right Intersection:

$$(X \not\twoheadrightarrow Y|ZW) \;\&\; (X \not\twoheadrightarrow W|ZY) \implies (X \not\twoheadrightarrow YW|Z) \tag{11.22}$$

Left Intersection:

$$(X \not\twoheadrightarrow Y|ZW) \;\&\; (W \not\twoheadrightarrow Y|ZX) \implies (XW \not\twoheadrightarrow Y|Z) \tag{11.23}$$

where $(X \not\twoheadrightarrow Y|Z)_G$ has the interpretation that every directed path from X to Y in a directed graph G contains at least on node in Z.

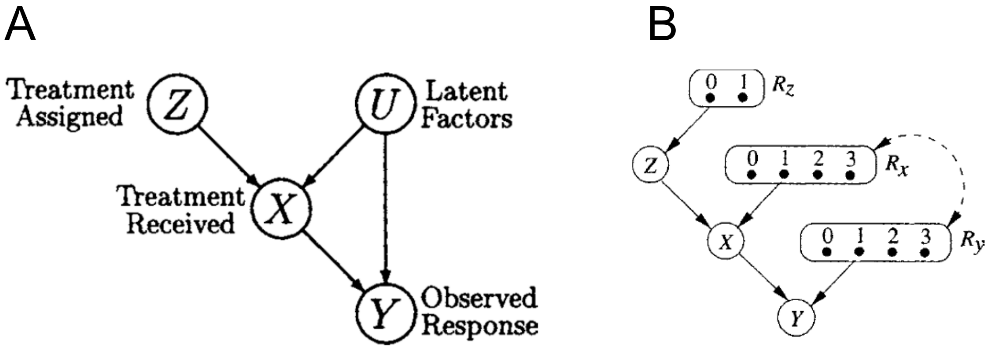


Figure 9. A randomized clinical trial with partial compliance and an instrumental variable. A. We cannot always conduct perfectly randomized clinical trials. We may bound causal effects by using an instrumental variable, such as treatment assigned. B. For binary X, Y, and Z, (B) is equivalent to (A), which uses finite-state response variables R_x, R_y, and R_z. R_x and R_y substitute U and may, in general, be dependent.

12. Imperfect experiments: bounding causal effects and counterfactuals

It is not always possible to conduct perfect randomized clinical trials, for e.g. ethical or pragmatic reasons. Sometimes, we must admit a latent common cause (U) of the treatment a patient actually receives (X) and the outcome for that patient (Y). We have seen above that, in such a graph $X \leftarrow U \rightarrow Y$, $P(y|do(x))$ is unidentifiable. However, we may consider an instrumental variable $Z \rightarrow X$, which may count as e.g. a physician's encouragement to participate in a particular way in the study, see Fig. 9A. If Z is assigned randomly, then we will be able to form bounds (as opposed to precise point-estimates) on $P(y|do(x))$.

12.1. Average causal effect

Let X, Y, and Z be binary random variables and the subscript i denote an outcome for the corresponding variable, where $i \in \{0, 1\}$.

Definition 12.1 (Average causal effect). *The average causal effect ACE($X \rightarrow Y$) is defined as*

$$\mathit{ACE}(X \rightarrow Y) = P(y_1|do(x_1)) - P(y_1|do(x_0)) \tag{12.1}$$

Using Rule 7.2 of do-calculus,

$$\text{ACE}(X \rightarrow Y) = \sum_u [P(y_1|x_1, u) - P(y_1|x_0, u)]P(u). \tag{12.2}$$

We seek to perform a constrained optimization of finding the highest and lowest values of Eq. (12.2), in terms of the observed probabilities $P(y, x|z_0)$ and $P(y, x|z_1)$.

12.2. Canonical partitions

No matter the form of $y = f(x, u)$, since Y and X are binary, we can always partition the domain of U into four equivalence classes

$$\begin{aligned} f_0 : y = 0, \quad & f_1 : y = x \\ f_1 : y = \neg x, \quad & f_2 : y = 1. \end{aligned} \tag{12.3}$$

Therefore for any $P(u)$ there exists a $P(r),\ r = 0, 1, 2, 3$ which is given by the total weight assigned to the equivalence class corresponding to f_r. Since X, Y and Z are all binary variables, the state space of U divides into $4 \times 2 + 4 \times 2 = 16$ equivalence classes, since U affects both X and Y. For convenience, we can split the state space into two four-valued variables R_x and R_y. We can then write, e.g.

$$x = f_X(z, r_x) = \begin{cases} x_0 & \text{if } r_x = 0; \\ x_0 & \text{if } r_x = 1 \text{ and } z = z_0; \\ x_1 & \text{if } r_x = 1 \text{ and } z = z_1; \\ x_1 & \text{if } r_x = 2 \text{ and } z = z_0; \\ x_0 & \text{if } r_x = 2 \text{ and } z = z_1; \\ x_1 & \text{if } r_x = 3. \end{cases} \tag{12.4}$$

We can call subjects with compliance behaviour $r_x = 0, 1, 2, 3$ as *never-taker*, *complier*, *defier*, and *always-taker*. We can define a similar mapping for $y = f_Y(x, r_y)$ with $r_y = 0, 1, 2, 3$ corresponding to *never-recover*, *helped*, *hurt*, and *always-recover* (see Fig. 9B). The counterfactual Y_{x_0} may be written as

$$Y_{x_0} = \begin{cases} y_1 & \text{if } r_y = 2 \text{ or } r_y = 3, \\ y_0 & \text{otherwise} \end{cases} \tag{12.5}$$

and similar for Y_{x_1}. Hence,

$$P(y_1|do(x_0)) = P(r_y = 2) + P(r_3 = 3). \tag{12.6}$$

Given the definition of ACE in Eq.(12.2), the conditional distribution $P(y, x|z)$, and the probabilistic constraints that $\sum_{x,y} P(x, y|z_0) = \sum_{x,y} P(x, y|z_1) = 1$, one may write down $ACE(X \to Y)$ as a linear programming problem, where we seek to minimize/maximize ACE given the constraints (for details see Section 8.2.3 of Pearl (2009)). The **natural bound** yields a less tight bound, which is simpler in its form (see Section 8.2.4 of Pearl (2009)). The width of the natural bounds is given by the rate of noncompliance: $P(x_1|z_0) + P(x_0|z_1)$. Natural bounds are optimal when no patient is a contrarian. Note that if the variables are continuous, then dichotomized variables can be defined.

12.3. Effect of treatment on the treated

ETT is useful when we want to know the impact of the treatment *on the treated*, for e.g. deciding whether to maintain/terminate a program. It is defined as

$$ETT(X \to Y) \coloneqq P(Y_{x_1} = y_1|x_1) - P(Y_{x_0} = y_1|x_1). \tag{12.7}$$

For the same DAG as Fig. 9, bounds may be derived through a similar means as the above (see (Pearl, 2009)). Under conditions of *no intrusion* (i.e. $P(x_1|z_0) = 0$), $ETT(X \to Y)$ can be identified precisely, and is

$$ETT(X \to Y) = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(x_1|z_1)}. \tag{12.8}$$

12.4. Test for instruments

Casting the estimation of $ACE(X \to Y)$ as a linear programming problem can allow us to derive tests for whether a variable Z is a plausible instrumental variable: i.e. a variable which is exogenous relative to $P(y|do(x))$. The following inequalities follow, by asserting that the upper bound of ACE is greater than the lower bound

$$\begin{aligned}
&P(y_0, x_0|z_0) + P(y_1, x_0|z_1) \leq 1, \\
&P(y_0, x_1|z_0) + P(y_1, x_1|z_1) \leq 1, \\
&P(y_1, x_0|z_0) + P(y_0, x_0|z_1) \leq 1, \\
&P(y_1, x_1|z_0) + P(y_0, x_1|z_1) \leq 1. \qquad (12.9)
\end{aligned}$$

Violation of any of these inequalities implies that Z cannot be an instrumental variable. This is not guaranteed to screen out all violations of exogeneity, but may still be useful. Generalizations for multivalued, and continuous Z or Y are

$$\max_x \sum_y [\max_z P(y, x|z)] \leq 1 \qquad (12.10)$$

$$\int_y \max_z [f(y|x, z)P(x|z)]dy \leq 1 \; \forall x \qquad (12.11)$$

respectively. These are called the **instrumental inequality**. Interestingly, for continuous X, Fig. 9 places no constraint on the observed density whatsoever.

13. Software

13.1. Tetrad

- Allows you to build causal graphs, a variety of parametric models (Bayesian causal graphs, SEMs,...), instantiate them with particular parametrizations
- Allows you to do model search, given data. Can be combined to create a ground truth, simulate, then do inference, all inside the same environment
- Implemented as a Java GUI
- Available for download here
- Manual here here

13.2. `causaleffect`

- Build causal graphs, and perform do-calculus to identify causal effects in terms of observational densities Tikka and Karvanen (2017), see Fig. 3
- Implemented in an `R` package, `causaleffect`
- Jupyter Notebook example here
- Does not return a causal effect in its simplest form

Acknowledgements

Thanks to Aidan Marshall for many useful conversations on the topic of causal inference.

References

Nayak, P. P., 1994 Causal approximations. Artificial Intelligence **70**: 277–334.

Pearl, J., 2009 *Causality*. Cambridge university press.

Ramsey, J., M. Glymour, R. Sanchez-Romero, and C. Glymour, 2017 A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International journal of data science and analytics **3**: 121–129.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, 2006 A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7**.

Simon, H. A., 1977 Causal ordering and identifiability. In *Models of Discovery*, pp. 53–80, Springer.

Spirtes, P., C. Glymour, R. Scheines, and R. Tillman, 2010 Automated search for causal relations: Theory and practice. Carnegie Mellon University .

Tian, J. and J. Pearl, 2002 A general identification condition for causal effects. In *Aaai/iaai*, pp. 567–573.

Tikka, S. and J. Karvanen, 2017 Identifying causal effects with the R package causaleffect. J. Stat. Softw .

Verma, T. and J. Pearl, 1988 *Influence diagrams and d-separation*. UCLA, Computer Science Department.