Causal inference cheat sheet

Author: Juvid Aryaman
Last compiled: November 21, 2020

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 X on Y: $r_{XY} = \rho_{XY}\sigma_X/\sigma_Y = \sigma_{XY}/(\sigma_Y^2)$ (for the equation $X = r_{XY}Y + c + \mathcal{N}(0, \sigma^2)$)
- Conditional independence: $(X \perp\!\!\!\perp Y|Z) \iff P(x|y, z) = P(x|z)$

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

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa_i) \tag{1.1}$$

1.1. d-separation (blocking) in Bayesian networks

A path p is d-separated (or blocked) by a set of notes 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

where an arrow $pa_j \to x_j$ denotes 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.

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).

2. Functional causal models

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

$$x_i = f_i(pa_i, u_i), \quad i = 1, ..., n \tag{2.1}$$

where pa_i are the set of variables (parents) that directly determine the value of X_i and U_i represents errors (or "disturbances") due to omitted factors. 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

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, ..., n \tag{2.2}$$

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

2.1. 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):

$$x = u_1 \tag{2.3}$$

$$y = u_2 \tag{2.4}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.5}$$

Model 2 (treatment has an effect):

$$x = u_1 \tag{2.6}$$

$$y = xu_2 + (1 - x)(1 - u_2) \tag{2.7}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.8}$$

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.

2.2. 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 $Y = y$.

3. Causal Bayesian networks

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.(2.1) 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_i are removed. **Amputation is the difference between seeing and doing**.

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

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{3.1}$$

The $j \neq i$ denotes the removal of the term $P(x_i|pa_i)$ from Eq.(1.1) (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.(1.1).

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

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} P(x_1, ..., x_n|x_i', pa_i)P(pa_i) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{3.2}$$

Marginalization of the above leads to the following theorem.

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

$$P(y|\hat{x}_i') = \sum_{pa_i} P(y|x_i', pa_i)P(pa_i) \tag{3.3}$$

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

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 3.2.5: 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. *all parents of the cause are necessary to estimate the causal effect.*

4. Inferring causal structure

- IC algorithm is for inferring causal structure given observational data when there are no latent variables
- IC* algorithm is for inferring causal structure given observational data when there are latent variables. The PC algorithm is apparently more contemporary (see Spirtes et al 2010)
- 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 $Z_1 \to X \to Y \leftarrow Z_2$
- NOTEARS (Zheng et al. 2018) casts the structure learning problem as a continuous optimization problem over real matrices to avoid the superexponential combinatorial explosion with number of variables.

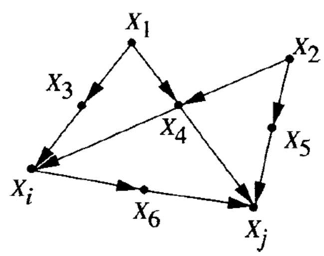


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 1. Example of the back-door criterion

5. 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.

5.1. The back-door criterion

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

Back-door 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 $X_j \in Y$.

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

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z). \tag{5.1}$$

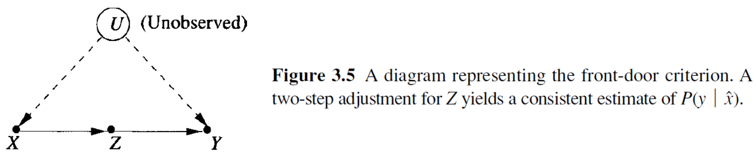
This corresponds to partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X on Y in each homogeneous group, and then averaging the results. Conditioning in this way means that the observation $X = x$ cannot be distinguished from an intervention $do(x)$.

5.2. The front-door criterion

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

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.



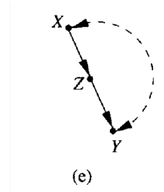


Figure 2. (Left) Example of the front-door criterion. The path $X \leftarrow U \rightarrow Y$ denotes an unobserved (latent) unobserved common cause. (Right) This is often represented as a **bi-directed path**.

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

$$P(y|\hat{x}) = \sum_{z} P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{5.2}$$

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 6 for a more general set of conditions).

6. 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**.

6.1. 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* $Z = z$

6.2. Rules

Rule 1 (Insertion/deletion of observations)

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{6.1}$$

This rule is a reaffirmation of d-separation (Section 1.1) 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 2 (Action/observation exchange)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{6.2}$$

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.

This rule 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.

Rule 3 (Insertion/deletion of actions)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}, \overline{Z(W)}}} \tag{6.3}$$

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

This rule 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.

6.3. 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 is there exists a finite sequence of transformations conforming to one of the three rules in Section 6.2 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 4.3.3 of Causality (this is presumably what DoWhy uses).

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).

- **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. 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. 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)

7.1. 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 $P^*(x|z)$.

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

$$\begin{aligned} P(y|do(X = g(z))) &= \sum_z P(y|do(X = g(z)), z)P(z|do(X = g(z))) \qquad (7.1) \\ &= \sum_z P(y|\hat{x}, z)|_{x=g(z)}P(z) \\ &= E_z[P(y|\hat{x}, z)|_{x=g(z)}]. \end{aligned}$$

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,

$$P(y)|_{P^*(x|z)} = \sum_x \sum_z P(y|\hat{x}, z)P^*(x|z)P(z). \qquad (7.2)$$

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.

7.2. 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 7.1. 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:*

$$Z_k \subseteq N_k \tag{7.3}$$

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

$$(Y \perp\!\!\!\perp X_k | X_1, ..., X_{k-1}, Z_1, ..., Z_k)_{G_{\underline{X}_k, \overline{X}_{k+1}, ..., \overline{X}_n}} \tag{7.4}$$

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

$$P(y|\hat{x}_1, ..., \hat{x}_n) = \sum_{z_1, ..., z_n} P(y|z_1, ..., z_n, x_1, ..., x_n) \times \prod_{k=1}^{n} P(z_k|z_1, ..., z_{k-1}, x_1, ..., x_{k-1}) \tag{7.5}$$