Chapter 2

Treatment Effect Identification and Estimation

2.1 Motivation

A large portion of research in causality is concerned with what may be broadly called *treatment effect estimation*. Treatment effect estimation is concerned with determining the effect of an action, or a sequence of actions, on a particular outcome or set of outcomes. For example, in economics, we might ask:

"What is the average increase in lifetime earnings for individuals with a bachelors degree compared to only high school education?"

In healthcare, we might ask:

"For a patient with kidney stones, does surgery or medicine have a higher success rate in eliminating symptoms after one month?"

Abstractly, these questions involve estimating an **average treatment effect**

$$\mathbb{E}[Y \mid \texttt{do}(A = 1)] - \mathbb{E}[Y \mid \texttt{do}(A = 0)],$$

the difference in the average outcome Y of two interventions, $\texttt{do}(A = 1)$ and $\texttt{do}(A = 0)$.

Such questions may be extended in a variety of ways. First, we may care about continuous-valued treatments. For example, in economics, we might ask:

"What is the average increase in GDP for x billion dollars of investment in infrastructure?"

In healthcare, we might ask:

"What is the average reduction in tumor size for x milligrams of a PD-1 inhibitor (a common immunotherapy drug for breast cancer)?"

Abstractly, these questions involve estimating a **dose-response curve**, i.e. the function

$$f(x) = \mathbb{E}[Y \mid \texttt{do}(A = x)].$$

Second, we may care not only about population averages, but about averages for specific subpopulations. For example, in economics, we might ask:

"Given the annual income of an individual's parents, what is the average increase in lifetime earnings for each additional year of college?"

In healthcare, we might ask:

"Given the size of an individual's kidney stone, does surgery or medicine have a higher success rate?"

Abstractly, these questions involve estimating a **conditional average treatment effect**

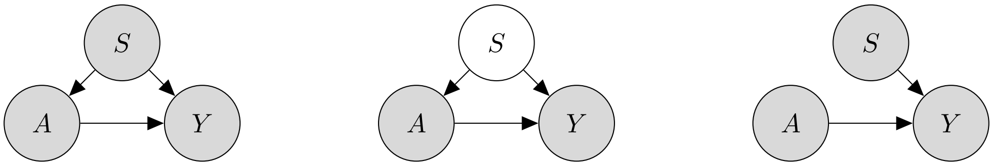
$$f(\mathbf{s}) = \mathbb{E}[Y \mid \mathbf{S} = \mathsf{s}, \mathtt{do}(A = 1)] - \mathbb{E}[Y \mid \mathbf{S} = \mathsf{s}, \mathtt{do}(A = 0)]$$

The subfield of causality devoted to answering such questions may also be called **policy evaluation** or **causal inference**. This is a very rich subfield, and we will only scratch its surface.

2.1.1 Running Example: Kidney Stone Treatment

In this lecture, we will focus on strategies for estimating the average treatment effect (ATE). To motivate why this is both important and non-trivial, we will hone in on the question introduced above:

"For a patient with kidney stones, does surgery or medicine have a higher success rate in eliminating symptoms after one month?"



(a) Stone size S is observed. (b) Stone size S is unobserved. (c) Intervention $\mathtt{do}(A = a)$.

Figure 2.1: DAGs representing the causal relationships between treatment A, kidney stone size S, and outcome Y. DAGs **(a)** and **(b)** represent the observed data-generating process. DAG **(c)** represents a **do**-intervention applied to the treatment variable A.

Example 2.1. *Let $S \in \{0, 1\}$ denote the size of a patient's kidney stone, where $S = 1$ denotes that the patient has large kidney stones and $S = 0$ denotes that they have a regular kidney stones. Let $A \in \{0, 1\}$ denote the patient's treatment, where $A = 1$ denotes that the patient receives the more intensive treatment (surgery) and $A = 0$ denotes that the patient receives the less intensive treatment (medication). Let $Y \in \{0, 1\}$ denote whether the kidney stones are eliminated after one month, where $Y = 1$ denotes that they are eliminated and $Y = 0$ denotes that they are not. Let the causal graph $\mathcal{G}$ be the DAG in Figure 2.1a. Assume that the data are generated according to the structural causal model M:*

$$S = \mathbb{1}_{\varepsilon_1 \leq 0.5} \qquad \varepsilon_1 \sim \mathsf{Unif}([0, 1])$$

$$A = \begin{cases} \mathbb{1}_{\varepsilon_2 \leq 0.1} & \textit{if } S = 0 \\ \mathbb{1}_{\varepsilon_2 \leq 0.9} & \textit{if } S = 1 \end{cases} \qquad \varepsilon_2 \sim \mathsf{Unif}([0, 1])$$

$$Y = \begin{cases} \mathbb{1}_{\varepsilon_3 \leq 0.7} & \textit{if } S = 0, A = 0 \\ \mathbb{1}_{\varepsilon_3 \leq 0.9} & \textit{if } S = 0, A = 1 \\ \mathbb{1}_{\varepsilon_3 \leq 0.1} & \textit{if } S = 1, A = 0 \\ \mathbb{1}_{\varepsilon_3 \leq 0.3} & \textit{if } S = 1, A = 1 \end{cases} \qquad \varepsilon_3 \sim \mathsf{Unif}([0, 1])$$

where the distributions of the exogenous variables $\varepsilon_1, \ldots, \varepsilon_3$ are mutually independent.

In the example, there is a 50% chance that a patient has large vs regular-sized kidney stones. The chosen treatment A is affected by the size of a patient's kidney stones: having larger stones results in a higher probability of surgery (90%) compared to having smaller stones (10%). Finally, the patient's outcome is influenced by both kidney stone size and the chosen treatment: having smaller kidney stones increases the patient's likelihood of recovery, as does being treated with surgery compared to medication.

2.1.2 Why Causality?

Comparing the relative treatment effects of surgery vs medication on kidney stone removal is an **interventional** question (rung 2 on the ladder of causation). It cannot be answered by using only standard statistical reasoning (rung 1). To see why, we will now show that the expected treatment effect is **not** equal to the expected patient outcome given that we *observe* that the treatment was applied, i.e., $\mathbb{E}[Y \mid \mathtt{do}(A = a)] \neq \mathbb{E}[Y \mid A = a]$.

Example 2.2. *We consider the* ***statistical*** *relationship between kidney stone treatment and patient outcomes under the SCM M given in Example 2.1. The expected patient outcome Y given that we observe that that patient was given treatment A = a can be computed as follows:*

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid A = a] &= \mathbb{P}_{\mathcal{X}}(Y = 1 \mid A = a) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1, S = s | A = a) \quad (\textit{law of total probability}) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s | A = a) \quad (\textit{multiplication rule}) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\frac{\mathbb{P}_{\mathcal{X}}(A = a | S = s)\mathbb{P}_{\mathcal{X}}(S = s)}{\mathbb{P}_{\mathcal{X}}(A = a)} \quad (\textit{Bayes' rule}) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\frac{\mathbb{P}_{\mathcal{X}}(A = a | S = s)\mathbb{P}_{\mathcal{X}}(S = s)}{\sum_{s'} \mathbb{P}_{\mathcal{X}}(A = a | S = s')\mathbb{P}_{\mathcal{X}}(S = s')} \quad (\textit{law of total probability}) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\frac{\mathbb{P}_{\mathcal{X}}(A = a | S = s)}{\sum_{s'} \mathbb{P}_{\mathcal{X}}(A = a | S = s')} \quad (\forall s\ \mathbb{P}_{\mathcal{X}}(S = s) = 0.5,\ \textit{so terms cancel}) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\frac{\mathbb{P}_{\mathcal{X}}(A = a | S = s)}{\mathbb{P}_{\mathcal{X}}(A = a | S = 0) + \mathbb{P}_{\mathcal{X}}(A = a | S = 1)} \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\frac{\mathbb{P}_{\mathcal{X}}(A = a | S = s)}{0.9 + 0.1} \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = a, S = s)\mathbb{P}_{\mathcal{X}}(A = a | S = s)
\end{aligned}$$

Using this formula, we can compute the expected patient outcome Y given that we observe that the patient had surgery (A = 1):

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid A = 1] &= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 1, S = s)\mathbb{P}_{\mathcal{X}}(A = 1 \mid S = s) \\
&= \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 1, S = 0)\mathbb{P}_{\mathcal{X}}(A = 1 \mid S = 0) + \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 1, S = 1)\mathbb{P}_{\mathcal{X}}(A = 1 \mid S = 1) \\
&= (0.9 * 0.1 + 0.3 * 0.9) = 0.36
\end{aligned}$$

We can similarly compute the expected outcome Y given that we observe that the patient was given medication (A = 0):

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid A = 0] &= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 0, S = s)\mathbb{P}_{\mathcal{X}}(A = 0 \mid S = s) \\
&= \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 0, S = 0)\mathbb{P}_{\mathcal{X}}(A = 0 \mid S = 0) + \mathbb{P}_{\mathcal{X}}(Y = 1 | A = 0, S = 1)\mathbb{P}_{\mathcal{X}}(A = 0 \mid S = 1) \\
&= (0.7 * 0.9 + 0.1 * 0.1) = 0.64
\end{aligned}$$

Overall, using statistical reasoning to compare patient outcomes leads us to conclude that medication is the more effective treatment, since it is associated with a higher success rate. However, as we will see next, this conclusion is flawed, and incorporating causal reasoning will allow us to arrive at the correct one.

Example 2.3. *We consider the* ***causal*** *relationship between kidney stone treatment and patient outcomes under the SCM M given in Example 2.1. In particular, we are interested in the expected value of a patient's outcome Y when we apply a intervention to the treatment variable A to set it to a particular value $A = a$, i.e., $\mathbb{E}_{\mathcal{X}}[Y \mid \texttt{do}(A = a)]$. In order to obtain this, we need to reason over the interventional SCM M^I, which is an SCM with:*

- *The same exogenous distribution as M*
- *The same causal mechanisms as M for non-intervened variables (i.e, the causal mechanisms of S and Y stay the same)*
- *The causal mechanism for A set to $A = a$*

The associated DAG is shown in Figure 2.1c. The absence of an arrow from S to A reflects that in the interventional SCM M^I, the causal mechanism of A no longer depends on the value of S. Recall from Section 1.6 that the interventional distribution $\mathbb{P}_{\mathcal{X}}(Y \mid \texttt{do}(A = a))$ of the SCM M is the entailed distribution of the interventional SCM M^I. With this in mind, we can compute the expected treatment effect as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid \texttt{do}(A = a)] &= \mathbb{P}_{\mathcal{X}}(Y = 1 \mid \texttt{do}(A = a)) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1, S = s|\texttt{do}(A = a)) \quad \textit{(law of total probability)} \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1|\texttt{do}(A = a), S = s)\mathbb{P}_{\mathcal{X}}(S = s|\texttt{do}(A = a)) \quad \textit{(multiplication rule)} \\
&\textit{(}A = a \textit{ with probability } 1 \textit{ in } M^I\textit{, so we can condition on } A\textit{)} \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1|A = a, \texttt{do}(A = a), S = s)\mathbb{P}_{\mathcal{X}}(S = s|\texttt{do}(A = a)) \\
&= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1|A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s) \quad \textit{(consistency)}
\end{aligned}$$

For the last step, we used the property of ***consistency*** *from Section 1.6: if a variable X is* ***not*** *intervened on, then the entailed conditional distribution of X given its parents is the same in both the original and interventional SCMs, i.e., $\mathbb{P}_{\mathcal{X}}(X \mid \mathrm{pa}_{\mathcal{G}}(X), \texttt{do}(A = a)) = \mathbb{P}_{\mathcal{X}}(X \mid \mathrm{pa}_{\mathcal{G}}(X))$.*

Using this formula, we can compute the expected value of a patient's outcome Y when surgery is applied as the treatment as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid \texttt{do}(A = 1)] &= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1|A = 1, S = s)\mathbb{P}_{\mathcal{X}}(S = s) \\
&= \mathbb{P}_{\mathcal{X}}(Y = 1|A = 1, S = 0)\mathbb{P}_{\mathcal{X}}(S = 0) + \mathbb{P}_{\mathcal{X}}(Y = 1|A = 1, S = 1)\mathbb{P}_{\mathcal{X}}(S = 1) \\
&= 0.9 * 0.5 + 0.3 * 0.5 = 0.6
\end{aligned}$$

We can similarly compute the expected value of a patient's outcome Y when medication is applied as the treatment as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[Y \mid \texttt{do}(A = 0)] &= \sum_s \mathbb{P}_{\mathcal{X}}(Y = 1|A = 0, S = s)\mathbb{P}_{\mathcal{X}}(S = s) \\
&= \mathbb{P}_{\mathcal{X}}(Y = 1|A = 0, S = 0)\mathbb{P}_{\mathcal{X}}(S = 0) + \mathbb{P}_{\mathcal{X}}(Y = 1|A = 0, S = 1)\mathbb{P}_{\mathcal{X}}(S = 1) \\
&= 0.7 * 0.5 + 0.1 * 0.5 = 0.4
\end{aligned}$$

Finally, we can compute the ATE as: $\mathbb{E}_{\mathcal{X}}[Y \mid \mathtt{do}(A=1)] - \mathbb{E}_{\mathcal{X}}[Y \mid \mathtt{do}(A=0)] = 0.6 - 0.4 = 0.2$. *Thus, we see that using causal reasoning leads to the correct conclusion that surgery is the more effective treatment – the opposite of the conclusion we reached when analyzing purely statistical relationships.*

Remark 2.1. *The kidney stone treatment example is an example of a phenomenon known as* ***Simpson's Paradox****. This occurs when a trend appears within groups (i.e., surgery is associated with better outcomes within both the large- and regular-sized kidney stone groups), but reverses when the groups are combined (e.g., surgery is associated with worse outcomes on average). As we just saw, this paradox can be resolved by incorporating causal reasoning and accounting for confounding variables.*

2.1.3 Identifiability

Typically, when we are interested in determining the causal effect of a treatment, we do not know the underlying data generating process (i.e., the full SCM). Instead, we observe samples from the entailed distribution of the SCM. For example, in the kidney stone treatment case, it is unlikely that we would know the exact function mapping stone size S and treatment A to patient outcome Y. However, we can expect to observe samples from the joint distribution $\mathbb{P}_{\mathcal{X}}(Y, S, A)$.

The study of treatment effect estimation focuses on the case in which we have access to entailed distribution $\mathbb{P}_{\mathcal{X}}(X)$ and the causal graph $\mathcal{G}$, but not the full SCM (i.e., not the causal mechanisms or the exogenous distribution). If we have access to the *experimental* data distribution (i.e., the entailed distribution of the interventional SCM), then computing treatment effects is straightforward. However, if we instead have access to the *observational* data distribution (i.e., the entailed distribution of the SCM where the treatment variable has not been manipulated), it is not always possible to identify treatment effects.

Experimental data. When a randomized controlled trial (RCT) can be conducted, questions about treatment effects are often conceptually easy to answer. In a randomized controlled trial, patients are randomly assigned to treatment ($\mathtt{do}(A=1)$) or control ($\mathtt{do}(A=0)$). Thus, one obtains samples from the distributions $\mathbb{P}_{\mathcal{X}}(\mathcal{X} \mid \mathtt{do}(A=1))$ and $\mathbb{P}_{\mathcal{X}}(\mathcal{X} \mid \mathtt{do}(A=0))$, from which conditional expectations can be directly estimated using averaging.

However, even in the study of randomized controlled trials, there is a rich set of questions, for example, questions about experimental design (i.e., the assignment of patients to treatment and control) or about the use of auxiliary variables.

Identifiability from observational data. In this the remainder of this lecture, we will discuss how to answer questions about treatment effects from only observational data, i.e., data where the treatment variable has not been manipulated. This is of great interest in many fields where randomized controlled trials are infeasible due to cost or ethical concerns. For instance, in healthcare, there is a large amount of data in electronic health records (EHRs) about patients, treatments, and outcomes, which does not come from RCTs. We will now formally define the issue of identifiability of an interventional distribution from observational data.

Definition 2.1. *Let* $\mathcal{G}$ *be an DAG. We say that an interventional distribution* $\mathbb{P}(Y \mid \mathtt{do}(\mathbf{A}=\mathbf{a}))$ *is (observationally)* **identifiable** *from* $\mathcal{G}$ *if, for any two SCMs* M_a *and* M_b *with causal graph* $\mathcal{G}$ *and entailed distribution* $\mathbb{P}_{\mathcal{X}}(\mathcal{X})$*, we have* $\mathbb{P}_{\mathcal{X}}^{M_a}(Y \mid \mathtt{do}(\mathbf{A}=\mathbf{a})) = \mathbb{P}_{\mathcal{X}}^{M_b}(Y \mid \mathtt{do}(\mathbf{A}=\mathbf{a}))$.

We will now return to the kidney stone treatment example to show how an interventional distribution may not be identifiable. This occurs due to the presence of unobserved variables (e.g., unobserved confounding), so we will examine the case in which we do **not** observe kidney stone size S.

Example 2.4. *Let* $\mathcal{G}$ *be the DAG in Figure 2.1b. Let SCM* M *be as given in Example 2.1. Let* M' *be the* *structural causal model:*

$$\begin{aligned}
S &= \mathbb{1}_{\varepsilon_1 \leq 0.5} && \varepsilon_1 \sim \mathsf{Unif}([0,1]) \\
A &= \mathbb{1}_{\varepsilon_2 \leq 0.5} && \varepsilon_2 \sim \mathsf{Unif}([0,1]) \\
Y &= \begin{cases} \mathbb{1}_{\varepsilon_3 \leq 0.64} & \text{if } A = 0 \\ \mathbb{1}_{\varepsilon_3 \leq 0.36} & \text{if } A = 1 \end{cases} && \varepsilon_3 \sim \mathsf{Unif}([0,1])
\end{aligned}$$

where the distributions of the exogenous variables $\varepsilon_1, \ldots, \varepsilon_3$ are mutually independent.

Both models M and M' are consistent with causal graph $\mathcal{G}$[1] *and have the same entailed distribution $\mathbb{P}^M_{\mathcal{X}}(A,Y) = \mathbb{P}^{M'}_{\mathcal{X}}(A,Y)$. To see this, note that for both M and M':*

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}}(A,Y) &= \mathbb{P}_{\mathcal{X}}(Y|A)\mathbb{P}_{\mathcal{X}}(A) \\
&= \mathbb{P}_{\mathcal{X}}(Y|A) * 0.5 \\
&= \begin{cases} 0.36 * 0.5 = 0.18 & \text{if } Y = 0, A = 0 \\ 0.64 * 0.5 = 0.32 & \text{if } Y = 0, A = 1 \\ 0.64 * 0.5 = 0.32 & \text{if } Y = 1, A = 0 \\ 0.36 * 0.5 = 0.18 & \text{if } Y = 1, A = 1 \end{cases}
\end{aligned}$$

Despite having the same observational distributions, the interventional distributions of M and M' differ. $\mathbb{P}^{M'}_{\mathcal{X}}(Y = 1 \mid \mathtt{do}(A = 0)) = 0.64$ (since their is no dependence on S, we can read this directly from the causal mechanism of Y). However, $\mathbb{P}^{M}_{\mathcal{X}}(Y = 1 \mid \mathtt{do}(A = 0)) = 0.60$, as we showed in Example 2.3.

In contrast, if the variable confounding A and Y is observed (i.e., we assume access to $\mathbb{P}_{\mathcal{X}}(A, Y, S)$), then we may identify $\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a))$, as follows.

Theorem 2.1. *Assume that $\mathbb{P}_{\mathcal{X}}$ factorizes according to $\mathcal{G}$ in Figure 2.1a. Then*

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_s \mathbb{P}_{\mathcal{X}}(Y \mid A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s)$$

Proof. We have by the law of total probability that

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a), S = s)\mathbb{P}_{\mathcal{X}}(S = s \mid \mathtt{do}(A = a))$$

We may condition on $A = a$ without changing the probabilities, since $A = a$ with probability one:

$$= \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a), A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s \mid \mathtt{do}(A = a), A = a)$$

Since Y is not intervened and $\mathrm{pa}_{\mathcal{G}}(Y) = \{A, S\}$, we have by consistency that

$$= \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s \mid \mathtt{do}(A = a))$$

Since S is not intervened and $\mathrm{pa}_{\mathcal{G}}(S) = \varnothing$, we have by consistency that

$$= \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, S = s)\mathbb{P}_{\mathcal{X}}(S = s)$$

□

[1] In the SCM M' the variables A and Y do not depend on S. However, M' is still consistent with the causal graph $\mathcal{G}$. We can define equivalent causal mechanisms in which A and Y are trivial functions of S (e.g., add $S * 0$). The key point is that while we know that the causal effect direction is from S to A/Y, we don't know the strength of the effect (which could be zero).

Remark 2.2. *This formula should look familiar - it is what we ended up deriving to compute treatment effects in the kidney stone example (Example 2.3). Here we show that it holds more generally for causal graphs of the form shown in Figure 2.1a.*

Remark 2.3. *In general, we say that* $\mathbf{S}$ *is an* **adjustment set** *for* $\mathbb{P}(Y \mid \mathtt{do}(A = a))$ *if the following equation holds:*

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s}) \mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s})$$

In the following parts of the lecture, we will establish more general conditions under which an interventional distribution is identifiable from observational data.

2.2 Connecting DAGs and Probability Distributions

Before diving deeper into treatment effect identification formulas, we will first take a brief detour to provide some relevant background material. We will focus on establishing connections between DAGs and probability distributions.

2.2.1 Factorization

We first introduce a relationship between DAGs and probability distributions based on *factorization*.

Let $\mathcal{G}$ be a DAG on nodes $\mathcal{X}$ and $\mathbb{P}_{\mathcal{X}}$ be a distribution. We say that $\mathbb{P}_{\mathcal{X}}$ **factorizes** according to $\mathcal{G}$ if

$$\mathbb{P}_{\mathcal{X}}(\mathcal{X}) = \prod_{X_i \in \mathcal{S}} \mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i)) \tag{2.1}$$

Claim 2.1. *Let* $\mathbb{P}_{\mathcal{X}}$ *be the entailed distribution of a Markovian structural causal model with causal graph* $\mathcal{G}$*. Then* $\mathbb{P}_{\mathcal{X}}$ *factorizes according to* $\mathcal{G}$*.*

2.2.2 d-separation and Conditional Independence

Next, we introduce a second relationship between directed graphs and probability distributions based on d-separation and conditional independence. The definition of d-separation is fairly complex, so we will build up to it in smaller pieces.

As a first step, we consider paths in DAGs, and divide nodes on the path into two types.

Definition 2.2. *Let* $\gamma = \langle \gamma_1, \ldots, \gamma_M \rangle$ *be a path in DAG. We call a node* γ_m *on this path a* **collider** *on* γ *if* $\gamma_{m-1} \to \gamma_m \leftarrow \gamma_{m+1}$*, i.e., two arrowheads "collide" at* γ_m*. Otherwise, we call* γ_m *a* **non-collider***.*

Now, we define what it means for a path to be blocked at a certain node, with the conditions for being blocked differing for colliders and non-colliders.

Definition 2.3. *Given an DAG* $\mathcal{G}$*, a set* $\mathbf{S} \subseteq \mathcal{X}$*, and a path* γ*, we call a node* γ_m *on the path a:*

- **Blocked non-collider** *if* γ_m *is a non-collider and* $\gamma_m \in \mathbf{S}$*, or a*
- **Blocked collider** *if* γ_m *is a collider and* $\overline{\mathrm{de}}_{\mathcal{G}}(\gamma_m) \cap \mathbf{S} = \varnothing$*, i.e., neither* γ_m *nor any of its descendants belong to* $\mathbf{S}$*.*

Otherwise, we call the node γ_m **unblocked***.*

Now, we may define d-connection simply in terms of the path being completely unblocked.

Definition 2.4. *Let $\mathcal{G}$ be an DAG on nodes $\mathcal{X}$. Given two nodes X_i and X_j and a set $\mathbf{S} \subseteq \mathcal{X} \setminus \{X_i, X_j\}$, we call a path γ between X_i and X_j a* **d-connecting path** *if all nodes in γ are unblocked. We say that X_i and X_j are* **d-connected** *given $\mathbf{S}$ if there exists any d-connecting path.*

Conversely, we define d-separation as the situation where all paths are blocked given $\mathbf{S}$.

Definition 2.5. *Given an DAG $\mathcal{G}$, we say two nodes X_i and X_j are* **d-separated** *by a set $\mathbf{S}$ if they are not d-connected given $\mathbf{S}$. We denote this by $X_i \perp\!\!\!\perp_{\mathcal{G}} X_j \mid \mathbf{S}$. For disjoints sets $\mathbf{A}, \mathbf{B}$ and $\mathbf{S}$, we say $\mathbf{A}$ is d-separated from $\mathbf{B}$ given $\mathbf{S}$ if $X_i \perp\!\!\!\perp_{\mathcal{G}} X_j \mid \mathbf{S}$ for all $X_i \in \mathbf{A}$, $X_j \in \mathbf{S}$. We denote the complete set of d-separation statements in an DAG $\mathcal{G}$ as $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G})$, i.e.,*

$$\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) := \{(\mathbf{A}, \mathbf{B}, \mathbf{S}) : \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{S}\}$$

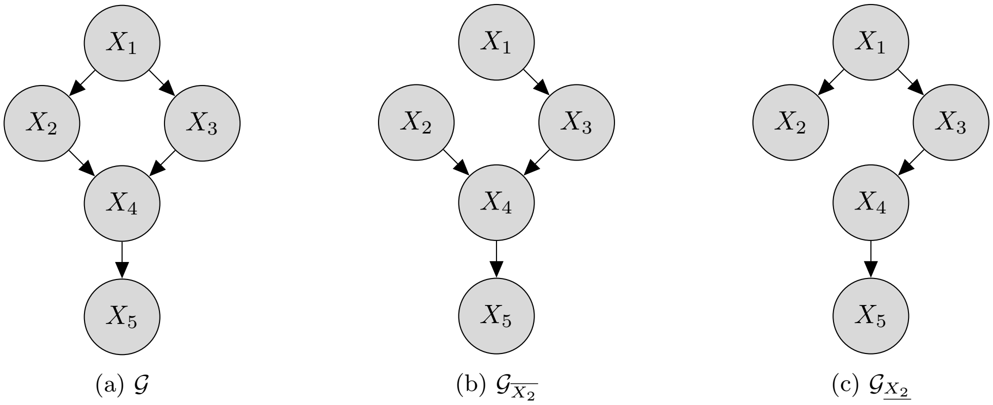


Figure 2.2: The DAG $\mathcal{G}$ (from Example 1.6 on genetics in mice) and augmented versions.

Example 2.5. *Let $\mathcal{G}$ be the DAG in Figure 2.2a.*

(a) *$\gamma = X_1 \to X_2 \to X_4$ is an d-connecting path from X_1 to X_4 given $\mathbf{S} = \varnothing$. This represents the fact that, knowing whether there was a genetic modification (X_1) tells us about the weight of Cheddar, through its effect on the weight of Mickey. However, γ is not an d-connecting path given $\mathbf{S} = \{X_2\}$, since X_2 is a blocked non-collider.*

(b) *$\gamma = X_2 \to X_4 \leftarrow X_3$ is not an d-connected path from X_2 to X_3 given $\mathbf{S} = \{X_1\}$, since X_4 is a blocked collider. Indeed, X_2 and X_3 are d-separated given X_1. This represents that, if we know whether there was a genetic modification, then knowing Mickey's weight tells us nothing about Minnie's weight.*

(c) *However, $\gamma = X_2 \to X_4 \leftarrow X_3$ is an d-connecting path given $\mathbf{S} = \{X_1, X_4\}$, since X_4 is unblocked. This represents the fact that, if we also know Cheddar's weight, then Mickey and Minnie's weights are again related. For example, suppose we know that Cheddar has a high weight. If we find that Minnie has a low weight, this means it is more likely that Mickey has a high weight to account for Cheddar's weight. This form of reasoning is commonly called* explaining away.

(c) *Similarly, $\gamma = X_2 \to X_4 \leftarrow X_3$ is an d-connecting path given $\mathbf{S} = \{X_1, X_5\}$: if we know Gouda's weight instead of Cheddar's, then Mickey and Minnie's weights are still related.*

We will now establish a connection between d-separation in DAGs and conditional independence in probability distributions.

Definition 2.6. *Given a distribution $\mathbb{P}_{\mathcal{X}}$, we define its* **independence model** *as the set of all conditional independence statements*

$$\mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_{\mathcal{X}}) := \{(\mathbf{A}, \mathbf{B}, \mathbf{S}) : \mathbf{A} \perp\!\!\!\perp_{\mathbb{P}_{\mathcal{X}}} \mathbf{B} \mid \mathbf{S}\}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{S}$ are disjoint subsets of $\mathcal{X}$.

Definition 2.7. *Let $\mathcal{G}$ be a DAG on nodes $\mathcal{X}$ and $\mathbb{P}_{\mathcal{X}}$ be a distribution. We say that $\mathbb{P}_{\mathcal{X}}$ is* **Markov** *with respect to $\mathcal{G}$ if $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) \subseteq \mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_{\mathcal{X}})$.*

Remark 2.4. *Note the direction of the inclusion: every d-separation in $\mathcal{G}$ implies a corresponding conditional independence in $\mathbb{P}_{\mathcal{X}}$. However, $\mathbb{P}_{\mathcal{X}}$ may include additional conditional independences, beyond the d-separations of $\mathcal{G}$. For example, if $\mathbb{P}_{\mathcal{X}}$ is a product distribution, then $\mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_{\mathcal{X}})$ includes all conditional independence statements, and thus $\mathbb{P}_{\mathcal{X}}$ is Markov with respect to any DAG $\mathcal{G}$.*

Requiring that $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) = \mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_{\mathcal{X}})$ is a much stronger condition which we call **faithfulness** *of $\mathbb{P}_{\mathcal{X}}$ to $\mathcal{G}$. We will return to the concept of faithfulness when we discuss structure learning.*

Theorem 2.2. *Let $\mathbb{P}_{\mathcal{X}}$ factorize according to a DAG $\mathcal{G}$. Then $\mathbb{P}_{\mathcal{X}}$ is Markov to $\mathcal{G}$, i.e., every d-separation in $\mathcal{G}$ entails a conditional independence in $\mathbb{P}_{\mathcal{X}}$.*

Remark 2.5. *For the sake of brevity, we will not prove this theorem. But if you're interested, check out the lecture notes from last year's version of this course (Lecture 2 notes). The converse of Theorem 2.2 also holds: if $\mathbb{P}_{\mathcal{X}}$ is Markov to $\mathcal{G}$, then $\mathbb{P}_{\mathcal{X}}$ factorizes according to $\mathcal{G}$ (although we will not need this fact). For a proof, see Theorem 3.27 of Lauritzen (1996).*

2.3 Non-Parametric Identification Formulas

We now return to considering formulas for identifying causal effects. In this section, we will discuss two well-known formulas that use the concept of d-separation. We will use the following result:

Theorem 2.3 (Action-Observation Exchange)**.** *Let I be an intervention with $\mathcal{X}(I) = \mathbf{A}$. Suppose $\zeta^I \perp\!\!\!\perp_{\mathcal{G}^I} Y \mid \mathbf{S}, \mathbf{A}$. Then*

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{S} = \mathbf{s}, \mathtt{do}(\mathbf{A} = \mathbf{a})) = \mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{S} = \mathbf{s}, \mathbf{A} = \mathbf{a}).$$

Proof. $\mathbb{P}_{\mathcal{X}^I}$ factorizes according to $\mathcal{G}^I$. By Claim 1.2,

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{S} = \mathbf{s}, \mathtt{do}(\mathbf{A} = \mathbf{a})) = \mathbb{P}_{\mathcal{X}^I}(Y \mid \mathbf{S} = \mathbf{s}, \zeta^I = 1)$$

By consistency and Theorem 2.2, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{X}^I}(Y \mid \mathbf{S} = \mathbf{s}, \zeta^I = 1) &= \mathbb{P}_{\mathcal{X}^I}(Y \mid \mathbf{S} = \mathbf{s}, \mathbf{A} = \mathbf{a}, \zeta^I = 1) \\ &= \mathbb{P}_{\mathcal{X}^I}(Y \mid \mathbf{S} = \mathbf{s}, \mathbf{A} = \mathbf{a}) \\ &= \mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{S} = \mathbf{s}, \mathbf{A} = \mathbf{a})\end{aligned}$$

Proving the result. □

Theorem 2.4. *Let I be an intervention with $\mathcal{X}(I) = \mathbf{A}$. Suppose $\mathbf{V} \perp\!\!\!\perp_{\mathcal{G}^I} \zeta^I \mid \mathbf{S}$. Then*

$$\mathbb{P}_{\mathcal{X}}(\mathbf{V} \mid \mathtt{do}(\mathbf{A} = \mathbf{a}), \mathbf{S}) = \mathbb{P}_{\mathcal{X}}(\mathbf{V} \mid \mathbf{S})$$

Proof. First, by Claim 1.2,

$$\mathbb{P}_{\mathcal{X}}(\mathbf{V} \mid \mathtt{do}(\mathbf{A} = \mathbf{a}), \mathbf{S} = \mathbf{s}) = \mathbb{P}_{\mathcal{X}^I}(\mathbf{V} \mid \zeta^I = 1, \mathbf{S} = \mathbf{s})$$

By Theorem 2.2, we have

$$\mathbb{P}_{\mathcal{X}^I}(\mathbf{V} \mid \zeta^I = 1, \mathbf{S} = \mathbf{s}) = \mathbb{P}_{\mathcal{X}}(\mathbf{V} \mid \mathbf{S} = \mathbf{s})$$

which gives the desired result. □

Definition 2.8. *We say that $\mathbf{S}$ satisfies the* **backdoor criterion** *for $\mathbb{P}(Y \mid \mathtt{do}(A = a))$ if $\zeta^I \perp\!\!\!\perp_{\mathcal{G}^I} Y \mid \mathbf{S}, A$ and $\mathbf{S} \perp\!\!\!\perp_{\mathcal{G}^I} \zeta^I$.*

Remark 2.6. *The backdoor criterion can be interpreted as follows:*

- $\zeta^I \perp\!\!\!\perp_{\mathcal{G}^I} Y \mid \mathbf{S}, A$ *says that* **S** *blocks all* **backdoor paths** *from A to Y, i.e., paths with arrows into both A and Y. All paths from ζ^I to Y with edges out of A will be blocked by conditioning on A, but conditioning on A unblocks paths with edges into A. The backdoor criterion requires that* **S** *blocks these paths.*
- $\mathbf{S} \perp\!\!\!\perp_{\mathcal{G}^I} \zeta^I$ *says that* **S** *contains no descendants of A, since ζ^I is only d-connected to A and its descendants.*

Theorem 2.5 (Backdoor Adjustment)**.** *Suppose that* **S** *satisfies the backdoor criterion for $\mathbb{P}(Y \mid \mathtt{do}(A = a))$. Then*

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s})$$

Proof. We have by the law of total probability that

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a), \mathbf{S} = \mathbf{s})\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s} \mid \mathtt{do}(A = a))$$

By Theorem 2.3, we have that

$$= \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s} \mid \mathtt{do}(A = a))$$

By Theorem 2.4,

$$= \sum_{\mathbf{s}} \mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s})$$

□

Example 2.6. *Let $\mathcal{G}$ be the graph in Figure 2.3. Sets that satisfy the backdoor criterion include $\{S_2, S_3\}$, $\{S_4, S_5\}$, and $\{S_1\}$, among others. However, adding S_6 or S_7 to any of these sets will result in a violation of the backdoor criterion, since S_6 and S_7 are d-connected to ζ^I.*

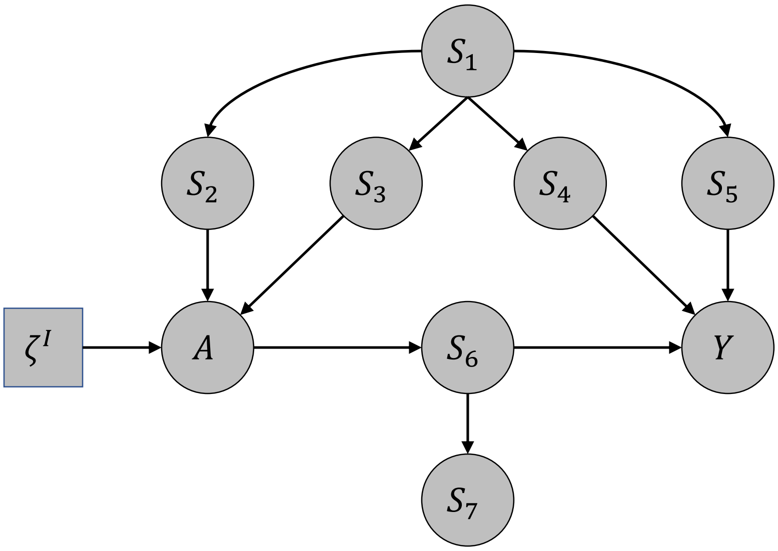


Figure 2.3: Graph for Example 2.6

Theorem 2.6 (Frontdoor Adjustment)**.** *Suppose that $\mathbb{P}_{\mathcal{X}}$ is the entailed distribution for a structural causal model with causal graph $\mathcal{G}$ in Figure 2.4. Then*

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{m}} \mathbb{P}_{\mathcal{X}}(\mathbf{M} = \mathbf{m} \mid A = a) \sum_{a'} (\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, A = a')\mathbb{P}_{\mathcal{X}}(A = a'))$$

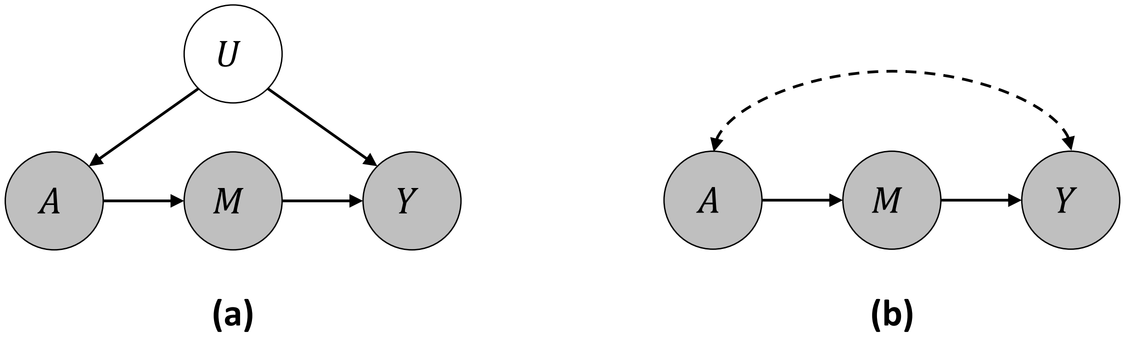


Figure 2.4: "Frontdoor" graph for Theorem 2.6

We first prove the following:

Claim 2.2.

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, \mathtt{do}(A = a)) = \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(\mathbf{M} = \mathbf{m})) \tag{2.2}$$

Proof. Let I_2 be an intervention on $\mathbf{M}$. Then, since $\zeta^{I_2} \perp\!\!\!\perp_{(\mathcal{G}^I)^{I_2}} Y \mid \mathbf{M}, A$, we have by Theorem 2.3 that

$$\begin{aligned}\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, \mathtt{do}(A = a)) &= \mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, A = a, \mathtt{do}(A = a)) \\ &= \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(\mathbf{M} = \mathbf{m}), A = a, \mathtt{do}(A = a)) \\ &= \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(\mathbf{M} = \mathbf{m}), \mathtt{do}(A = a))\end{aligned}$$

Since $Y \perp\!\!\!\perp_{(\mathcal{G}^I)^{I_2}} \zeta^I \mid \mathbf{M}$, we have by Theorem 2.4 that

$$= \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(\mathbf{M} = \mathbf{m}))$$

□

Proof of Theorem 2.7. We have by the law of total probability that

$$\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a)) = \sum_{\mathbf{m}} \mathbb{P}_{\mathcal{X}}(\mathbf{M} = \mathbf{m} \mid \mathtt{do}(A = a)) \mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, \mathtt{do}(A = a))$$

Since M is not intervened, by consistency we have

$$= \sum_{\mathbf{m}} \mathbb{P}_{\mathcal{X}}(\mathbf{M} = \mathbf{m} \mid A = a) \mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{M} = \mathbf{m}, \mathtt{do}(A = a))$$

Using Equation (2.2),

$$= \sum_{\mathbf{m}} \mathbb{P}_{\mathcal{X}}(\mathbf{M} = \mathbf{m} \mid A = a) \mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(\mathbf{M} = \mathbf{m}))$$

Now, A is a backdoor adjustment set for $\mathbb{P}(Y \mid \mathtt{do}(M = m))$, so we have the result. □

2.4 Parametric Identification Strategies

Previously, we considered identification where we make only *non-parametric* assumptions on our causal model: assumptions about the factorization of the distribution $\mathbb{P}_{\mathcal{X}}$, but not about any of the individual terms $\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$. In econometrics and other fields, it is common to achieve better identifiability

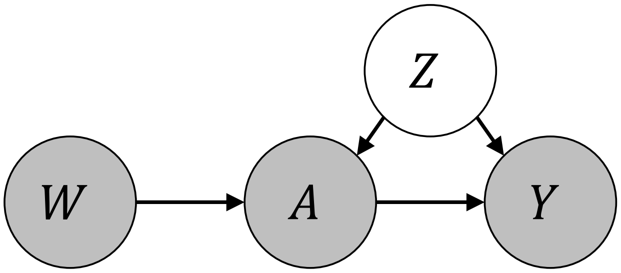


Figure 2.5: The instrumental variable model. W is called the **instrumental variable** for A.

guarantees by making *parametric* assumptions on our models. Such assumptions restrict the functional form of the causal mechanisms f_{X_i} to a class of functions that can be indexed by a finite-dimensional set of parameters. Common parametric models include linear models, where

$$f_{X_i}(\mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)) = \sum_{V_i \in \mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)} \beta_{V_i,X_i} V_i,$$

binary choice models, i.e.,

$$f_{X_i}(\mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)) = \mathbb{1}_{U_i \geq 0} \qquad \text{where } U_i = \sum_{V_i \in \mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)} \beta_{V_i,X_i} V_i,$$

exponential family models, and more. In the lecture, we will focus on linear models.

2.4.1 Instrumental variables

One of the most commonly used models for achieving identifiability via parametric assumptions is the **instrumental variable model**.

Definition 2.9. *We say that (W, A, Y) satisfy the* **instrumental variable model** *if there exists Z such that (Z, W, A, Y) are generated according to a linear structural causal model with causal graph $\mathcal{G}$ in Figure 3.1, i.e.,*

$$\begin{aligned} Z &= \varepsilon_z \\ W &= \varepsilon_w \\ A &= \beta_{za} Z + \beta_{wa} W + \varepsilon_a \\ Y &= \beta_{zy} Z + \beta_{ay} A + \varepsilon_y \end{aligned}$$

for $\varepsilon_z, \varepsilon_w, \varepsilon_a, \varepsilon_y$ independent. The variable W is called an **instrumental variable***.*

Remark 2.7. *An important feature of the instrumental variable model is that W does not have a direct effect on Y, i.e., all effects of W on Y are mediated via A. This feature is called the* **exclusion restriction***.*

Example 2.7. *A canonical example of the instrumental variable model is for the estimation of the effect of smoking (A) on cancer (Y). The association between smoking and higher rates of cancer has been well-known for over a century. However, in the 1950's, there was vigorous debate over whether smoking* caused *higher rates of cancer. Statisticians aligned with the tobacco industry noted that there may be some confounding factors (U), such as socioeconomic status, that explained the correlation between smoking and cancer. Possible instrumental variables for smoking include a tax on tobacco: higher taxes are likely to decrease the prevalence of smoking, but are not expected to have a direct effect on the rate of cancer.*

Theorem 2.7. *Let (W, A, Y) satisfy the instrumental variable model with $\beta_{wa} \neq 0$. Then β_{ay} is identifiable from $\mathbb{P}_{\mathcal{X}}(W, A, Y)$.*

Proof.

$$Y = (\beta_{zy} + \beta_{za}\beta_{ay})Z + \beta_{wa}\beta_{ay}W + \beta_{ay}\varepsilon_a + \varepsilon_y$$

We have $\mathbb{E}[A \mid W = \mathbf{w}] = \beta_{wa}\mathbf{w}$ by independence of W from Z and ε_a. Thus, β_{wa} is identifiable by linear regression of A on W.

Further,

$$Y = (\beta_{zy} + \beta_{za}\beta_{ay})Z + \beta_{wa}\beta_{ay}W + \beta_{ay}\varepsilon_a + \varepsilon_y$$

Thus, $\mathbb{E}[Y \mid W = \mathbf{w}] = \beta_{wa}\beta_{ay}\mathbf{w}$ by independence of W from Z, ε_a, and ε_y.

Thus, $\beta_{wa}\beta_{ay}$ is identifiable by linear regression of Y on W. Therefore, we can identify β_{ay} as the ratio of these two regression coefficients. □

Remark 2.8. *Note that our method of identifying β_{ay} uses two linear regressions. When the linear regression coefficients are estimated from data, the resulting estimator of β_{ay} is called the* **two-stage least squares** *estimator. Note that this estimator requires dividing by β_{wa}, which can lead to an estimator with high variance if the value of β_{wa} is near zero. In such cases, W is called a* **weak instrument***.*

2.5 Additional Reading

2.5.1 Non-Parametric Identification

- **The do-calculus and the ID algorithm**: Theorem 2.3 and Theorem 2.4 are special cases of the three rules of the **do-calculus** (Pearl, 1995). The do-calculus is a *complete* set of rules for determining identifiability from observational data: if a distribution $\mathbb{P}_{\mathcal{X}}(Y \mid \mathbf{S} = \mathbf{s}, \mathtt{do}(A = a))$ is identifiable, then the rules of the do-calculus can be applied to transform the interventional distribution into an expression which involves only observational distributions, called an **identification formula**. Shpitser and Pearl (2006) describes a complete algorithm to conduct this transformation: the algorithm either outputs an identification formula, or provides a certificate to show that the effect is not identifiable.

- **Counterfactual identification**: Many questions, especially those involving fairness, are best framed in terms of counterfactuals instead of interventions. Shpitser and Pearl (2008) provides a method for identifying counterfactual queries, and Malinsky et al. (2019) uses single world intervention graphs to define an analogue of the do-calculus called the potential outcome calculus or the **po-calculus**.

- **Identification from interventional data**: We have been considering identification from observational data alone. However, one may consider identifying the effect of an intervention from data where different interventions have taken place. This form of identification is considered for do-interventions in Lee et al. (2020) and for soft interventions in Correa and Bareinboim (2020).

- **Partial identification**: While an interventional query might not be identifiable, one may still be able to place *bounds* on it. This is particularly important if one wishes to determine whether a treatment effect is positive, i.e. one treatment does better than another. Deriving such bounds is called **partial identification**, see Richardson et al. (2014) for a review.

2.5.2 Parametric Identification

- **Nonlinear instrumental variables models and proxy variable models.** In general non-parametric models (those with no restrictions on the probability distributions $\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$), $\mathbb{P}_{\mathcal{X}}(Y \mid \mathtt{do}(A = a))$ is not identified for the causal graph in Figure 3.1. However, we do not need to go all the way to parametric models in order to guarantee identifiability: we may retain significant expressivity in these conditional distributions and still identify the interventional quantity, see e.g. Newey and Powell (2003) and Singh et al. (2019). Similar comments apply to the proxy variable model, see e.g. Kallus et al. (2021).

- **Identifiability in linear models.** There is a large literature on identifying causal effects in linear structural causal models, see for example Kumor et al. (2020), Barber et al. (2022), Drton et al. (2016). In linear models, identifiability can always be checked using techniques from computational algebra, but these are computationally intractable for large graphs. Thus, these papers seek to develop graphical conditions to check identifiability. Thus far, most papers focus on sufficient conditions for identifiability, though a necessary condition is given in Foygel et al. (2012). To the best of my knowledge, there is not yet a necessary and sufficient graphical characterization for identifiability.

- **Estimation of treatment effects.** In the past two lectures, we have been concerned with determining whether the interventional quantities of interest are identifiable from observational data. This is intrinsically a question about something that happens in the *infinite data* limit. This leaves open a major question: what should we do in practice, where we always have a finite amount of data?

  One intuitive answer is suggested by the observation that our proofs of identifiability often result in *identification formulas* which express the target interventional quantity in terms of observational quantities. This suggests that we could estimate the observational quantities from data (e.g., $\mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})$ and $\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s})$ in the context of backdoor adjustment), and then plug these estimates into our identification formula. This approach is called the **plug-in** approach, and works well in parametric models (including linear models or models where all variables are discrete). Indeed, if one estimates all parameters of a structural causal model using maximum likelihood estimation, then the plug-in approach can be shown to be (asymptotically) better than any other approach for estimating the target interventional quantity, using classical parametric efficiency theory.

  However, if one does not make such parametric assumptions, then the plug-in approach can be shown to be quite suboptimal. For example, if $\mathbf{S}$ is continuous-valued in the setting of backdoor adjustment, then $\mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})$ can be a complicated function which takes a large amount of data to estimate. The plug-in approach propagates errors in the estimation of these so-called **nuisance functions** to errors in the estimation of the target intervention quantity. More sophisticated methods can avoid this error propagation, see e.g. Kennedy (2022).

  Unfortunately, in this lecture series, we will not have time to properly address these statistical questions.

Bibliography

Barber, R. F., Drton, M., Sturma, N., and Weihs, L. (2022). Half-trek criterion for identifiability of latent variable models. *arXiv preprint arXiv:2201.04457*.

Correa, J. and Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100.

Drton, M. et al. (2016). Algebraic problems in structural equation modeling. *The 50th anniversary of Gröbner bases*, pages 35–86.

Foygel, R., Draisma, J., and Drton, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pages 1682–1713.

Kallus, N., Mao, X., and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Kumor, D., Cinelli, C., and Bareinboim, E. (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, pages 5501–5510. PMLR.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lee, S., Correa, J. D., and Bareinboim, E. (2020). General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pages 389–398. PMLR.

Malinsky, D., Shpitser, I., and Richardson, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. (2014). Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.