Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

March 5, 2025

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Causal Inference via Directed Acyclical Graphs and Nonlinear Structural Equation Models

7

"you are smarter than your data. Data do not understand causes and effects; humans do."

– Judea Pearl [1].

Here, we explore a fully nonlinear, nonparametric formulation of causal diagrams and their associated structural equation models (SEMs). These models offer a powerful and flexible tool for understanding the structures that underpin causal identification, allowing us to move beyond restrictive linear assumptions. Using these structures, we define potential outcomes—also known as counterfactuals—following what Judea Pearl terms the "First Law of Causal Inference," which establishes that SEMs naturally induce these outcomes. This foundation enables a systematic approach to causal analysis. Moreover, we can algorithmically verify, using the directed acyclic graphs (DAGs) that encode these structures, whether the conditional ignorability conditions necessary to transform predictive regressions into causal inferences are satisfied. In fact, given a DAG, we can derive sufficient adjustment sets—sets of variables to condition on in regressions—that enable us to uncover average causal effects. This process leverages the graphical representation of contextual knowledge to ensure that the statistical relationships we observe reflect true causal impacts, rather than mere correlations.

7.1 Introduction

The purpose of this module is to present a formal, fully nonlinear (nonparametric) formulation of structural equation models (SEMs) and their corresponding causal directed acyclic graphs (DAGs). We explore the concepts and identification results pioneered by Judea Pearl and his collaborators, as well as those developed by James M. Robins and his collaborators.[1]

SEMs define a recursive system of equations that generate variables. From these models, we can derive counterfactuals, also known as potential outcomes (POs).[2] We represent both factual and counterfactual variables using DAGs, leveraging their structure to deduce the conditional independence conditions (e.g., ignorability, exogeneity) required to transform predictive regressions into causal inferences.

We then examine two approaches for identifying variables to adjust for (condition on) when estimating the causal effect of a treatment on an outcome using DAGs: the backdoor adjustment approach and the counterfactual DAG approach. The backdoor criterion, developed by Judea Pearl, analyzes the factual DAG to identify a set of variables that blocks all backdoor paths – paths from the treatment to the outcome beginning with an arrow into the treatment– while ensuring these variables are not descendants of the treatment, thereby eliminating confounding influences. In contrast, the counterfactual DAG approach uses a modified DAG where the treatment is hypothetically fixed to a specific value, identifying a set of variables that "d-separates" the natural treatment value from the counterfactual outcome, ensuring conditional independence between the treatment and the counterfactual outcome given these variables.

This chapter is divided into two parts. The first introduces key concepts through a specific empirical example, while the second provides general, albeit more technical, mathematical definitions. Additional technical material is included in the appendix.

Notation. Consider a pair of random variables (or, equivalently, random vectors) U and V with joint probability (or mass) function $\mathsf{p}_{UV}(u, v)$ evaluated at (u, v). When no ambiguity arises, we denote $\mathsf{p}_{UV}(u, v)$ simply as $\mathsf{p}(u, v)$. Their marginal probability (or mass) functions are denoted by $\mathsf{p}_U(u)$ and $\mathsf{p}_V(v)$, or simply $\mathsf{p}(u)$ and $\mathsf{p}(v)$. We say that U and V are independent (denoted $U \perp\!\!\!\perp V$) if and only if the joint function factorizes as

$$\mathsf{p}(u, v) = \mathsf{p}(u)\,\mathsf{p}(v),$$

1: In 2011, J. Pearl received the A.M. Turing Award, the highest honor in Computer Science and Artificial Intelligence, "for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning." In his 1995 *Biometrika* article [2], Pearl frames his work as a generalization of the SEMs proposed by T. Haavelmo [3] in 1944 and others.

2: Pearl refers to the implication of POs from SEMs as the "First Law of Causal Inference."

or equivalently, if $\mathrm{E}[g(U)\ell(V)] = \mathrm{E}[g(U)]\mathrm{E}[\ell(V)]$ for any bounded functions g and ℓ. This definition of independence implies the ignorability or exclusion results,

$$\mathsf{p}(u \mid v) = \mathsf{p}(u), \quad \mathsf{p}(v \mid u) = \mathsf{p}(v),$$

which follow from Bayes' law. Conditional independence is defined similarly by replacing distributions and expectations with their conditional counterparts. Appendix 7.A reviews some useful results on conditional independence.

7.2 General DAG and SEMs via an Example

The best way to learn the main ideas behind modern SEMs and causal directed acyclic graphs DAGs is to work through a concrete example.

The Impact of 401(k) Eligibility on Financial Wealth

A 401(k) is a U.S. employer-sponsored retirement plan that allows workers to contribute a portion of their wages—often pre-tax—to investment accounts, sometimes with matching contributions from their employer. Figure 7.1 shows one possible causal diagram for this problem:

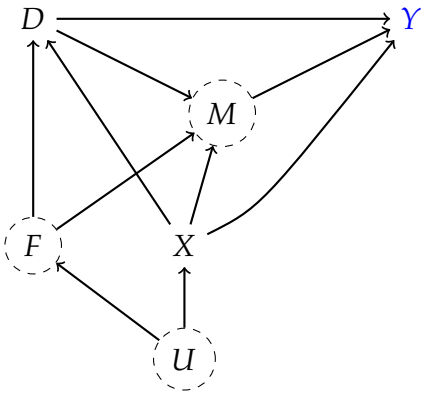


Figure 7.1: Causal Diagram for 401(k): Y represents net financial assets; D denotes eligibility for a 401(k) program; X includes observed worker-level covariates (e.g., income); F represents unobserved firm-level covariates; M denotes the employer's matching contribution; and U captures general latent factors.

This diagram represents how 401(k) eligibility (D) might affect an individual's net financial assets (Y) both directly and indirectly through the employer's matching contribution (M). It includes observed worker-level characteristics (X), unobserved firm-level characteristics (F), and latent factors (U) that may influence the pathway from eligibility to financial outcomes.

This representation reasonably captures the context of the underlying problem.[3]

3: Please check this assertion by querying AI.

The DAG as a Markovian Model

The DAG above formally represents the conditional dependencies among the variables, allowing us to express their joint distribution in terms of conditional distributions. In these graphs, each node represents a random variable (or vector), and an arrow from one node (a "parent") to another (a "child") indicates that the parent directly influences the child, establishing statistical dependency.

The *Markov property* states that each variable is conditionally independent of all non-parents (and non-descendants) given its parents. Consequently, the joint probability distribution can be written as the product of each variable's conditional distribution given its parents.

In our example, the variables are:

$$U, \quad F, \quad X, \quad D, \quad M, \quad Y,$$

with the following parent-child relationships based on the DAG:

- U has no parents (a "root" node),
- F has parent U,
- X has parent U,
- D has parents F and X,
- M has parents D, F, and X,
- Y has parents D, M, and X.

According to the chain rule for Markovian networks,[4] the joint distribution $\mathsf{p}(u, f, x, d, m, y)$ factorizes as:

$$\begin{aligned}\mathsf{p}(u, f, x, d, m, y) = \mathsf{p}(u) & \\ \times \mathsf{p}(f \mid u)\, \mathsf{p}(x \mid u) & \\ \times \mathsf{p}(d \mid f, x, u) & \\ \times \mathsf{p}(m \mid d, f, x) & \\ \times \mathsf{p}(y \mid d, m, x). &\end{aligned}$$

4: Markovian networks, also known as Bayesian Markovian Networks or simply Bayesian Networks, are a type of probabilistic graphical model. We use the term 'Markovian networks' as it is arguably more precise.

Because U is a root node, its distribution is simply $\mathsf{p}(u)$. Each subsequent variable is represented by its conditional distribution given its parents. This approach is a cornerstone of

probabilistic graphical models, clarifying the conditional independencies and potential pathways of influence among the variables.

The DAG as a Structural Equations Model

Following Pearl, we interpret the DAG as implying that (or being implied by) the following system of structural equations holds:

$$\begin{aligned} Y &:= f_Y(D, M, X, \epsilon_Y), \\ M &:= f_M(D, F, X, \epsilon_M), \\ D &:= f_D(F, X, \epsilon_D), \\ X &:= f_X(U, \epsilon_X), \\ F &:= f_F(U, \epsilon_F), \\ U &:= \epsilon_U, \end{aligned}$$

where

$$\epsilon_Y, \epsilon_M, \epsilon_D, \epsilon_X, \epsilon_F, \epsilon_U$$

are mutually independent stochastic shocks (which may be vector-valued), and f_Y, f_M, f_D, f_X, f_F are structural functions.

Here, each variable is defined as a function of its parent variables (as determined by the DAG) and its own exogenous noise ϵ. For instance, the equation

$$Y := f_Y(D, M, X, \epsilon_Y)$$

indicates that net financial assets Y are determined by eligibility D, the matching contribution M, observed covariates X, and an unobserved shock ϵ_Y. The assignment operator (:=) signifies that variables are generated recursively, starting from the root and proceeding through subsequent layers based on their parents and noise terms.

Intervention and Counterfactual DAG and SEM

Thus far, the DAG and SEM we have formulated lack inherent causal meaning. Causality emerges when we introduce the concept of an intervention. In particular, consider intervening by replacing D with a fixed value d in the equations for all descendants of D. By assumption, the structural equations remain invariant under such an intervention—this invariance

is the essence of their "structural" nature. Consequently, we obtain the counterfactual outcomes:

$$\begin{aligned} Y(d) &:= f_Y\big(d, M(d), X, \epsilon_Y\big), \\ M(d) &:= f_M\big(d, F, X, \epsilon_M\big), \end{aligned}$$

where $Y(d)$ denotes the potential net financial wealth under treatment d and $M(d)$ represents the matching contribution under d. The remaining equations remain unchanged. Thus, the complete counterfactual system is:

$$\begin{aligned} Y(d) &:= f_Y\big(d, M(d), X, \epsilon_Y\big), \\ M(d) &:= f_M\big(d, F, X, \epsilon_M\big), \\ D &:= f_D\big(F, X, \epsilon_D\big), \\ X &:= f_X\big(U, \epsilon_X\big), \\ F &:= f_F\big(U, \epsilon_F\big), \\ U &:= \epsilon_U. \end{aligned}$$

(Note: An alternative formulation, called the *do-counterfactual*, omits the equation for the naturally generated D; the version here is known as the *fix-counterfactual*.)

The invariance assumption ensures that the functional forms f_Y, f_M, etc., remain unchanged even when we set $D = d$ in the equations for Y and M. Although the original equation for D remains in the model, the intervention fixes D at d for the purpose of determining $Y(d)$ and $M(d)$.

We can now construct a counterfactual DAG—also known as a SWIG (Single World Intervention Graph)—that corresponds to this counterfactual system.

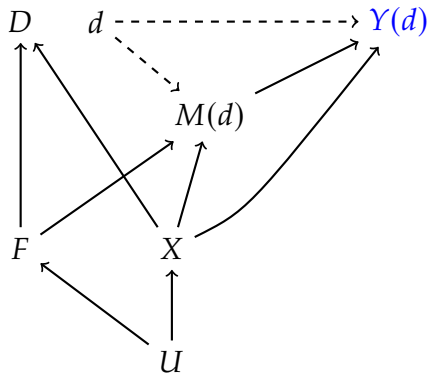


Figure 7.2: Counterfactual DAG for the intervention $D = d$. Here, $Y(d) = f_Y(d, M(d), X, \epsilon_Y)$ and $M(d) = f_M(d, F, X, \epsilon_M)$ reflect the intervention (with d shown as a deterministic node). The natural node D is still generated by $f_D(F, X, \epsilon_D)$, but its outgoing arrows are removed. Other nodes: X (worker-level covariates), F (firm-level covariates), and U (latent factors).

Conditional Ignorability/Exogeneity

The fact that the SEM implies potential outcomes is known as the First Law of Causal Inference. This equivalence means that nothing is lost by working with SEMs/DAGs instead of potential outcomes directly. Moreover, because SEMs/DAGs encapsulate the contextual knowledge of a problem, we can derive the conditional ignorability/exogeneity condition from the model rather than merely postulating it. For example, in our case we deduce that

$$Y(d) \perp\!\!\!\perp D \mid F, X,$$

which implies that

$$\mathrm{E}\big[Y(d) \mid F, X\big] = \mathrm{E}\big[Y \mid D = d, F, X\big],$$

allowing us to identify average causal (or treatment) effects by adjusting (or conditioning on F, X.)

There are two ways to verify that (F, X) satisfy this condition:

Functional (Structural) Argument.

In the counterfactual setting where we fix $D = d$, the relevant structural equations are

$$Y(d) = f_Y\big(d, M(d), X, \epsilon_Y\big) \quad \text{and} \quad M(d) = f_M\big(d, F, X, \epsilon_M\big).$$

The random variable D is still generated by

$$D = f_D\big(F, X, U, \epsilon_D\big).$$

Once we condition on F and X, the distribution of $Y(d)$ is determined solely by d, $M(d)$, X, and their associated noise terms, and is not influenced by the realized value of D. Formally, $Y(d)$ is statistically independent from D, conditional on F and X:

$$Y(d) \perp\!\!\!\perp D \mid F, X.$$

In other words, once F and X are given, knowing D adds no additional information about $Y(d)$.

A similar argument shows that D is not ignorable when conditioning on worker characteristics X alone:

$$Y(d) \not\perp\!\!\!\perp D \mid X.$$

Omitted firm characteristics F induce a dependency between the potential outcomes and the treatment D, posing a problem for studies that control for X but not F.[5]

5: In such cases, omitted variable bias can still be studied and bounded; see later chapters and [4].

D-Separation (Graphical) Argument

In the counterfactual DAG, $Y(d)$ receives inputs from $M(d)$, X, and the fixed node d. Although D remains in the graph (generated by its usual parents F, X, and U), there is no arrow from D to $Y(d)$. Any path from D to $Y(d)$ must traverse F or X. For example, the paths are:

1. $D \leftarrow X \rightarrow Y(d)$,
2. $D \leftarrow F \rightarrow M(d) \rightarrow Y(d)$,
3. $D \leftarrow F \leftarrow U \rightarrow X \rightarrow Y(d)$.

Conditioning on F and X is said to *block* these paths (conditioning on a node severs information flow), which then makes D to be *d-separated* from $Y(d)$ given $\{F, X\}$.[6] By the equivalence between d-separation and conditional independence,[7] we conclude that

$$Y(d) \perp\!\!\!\perp D \mid F, X.$$

6: See the formal definition of d-separation and blocking later in the chapter.

7: Equivalence of d-separtion and conditional independence is called Global Markov property and is a fundamental result in the DAG theory.

Backdoor Blocking (Graphical) Argument

Another approach for identifying the average causal effect of D on Y uses the original DAG instead of the counterfactual DAG. We note though that this principle was in fact derived by J. Pearl [2] from the counterfactual DAG of the form stated above.

The goal is to identify a set Z that *blocks* all *backdoor paths* between D and Y. A set Z satisfies the *backdoor criterion* if:

1. No variable in Z is a descendant of D, and
2. Z blocks[8] every backdoor path from D to Y (a backdoor path starts with an arrow into D).

8: See the formal definition of blocking later in the chapter.

The first rule prevents blocking causal paths from D to Y, such as $D \rightarrow M \rightarrow Y$.[9] The second rule ensures that conditioning on Z eliminates all non-causal paths that could confound the relationship between D and Y. Thus condition on Z, the statistical association between Y and D only reflects the causal channels.

9: It also prevents conditioning on colliders, examples of which we have seen the previous chapter.

In the 401(k) diagram, the backdoor paths from D to Y run through F and X:

1. $D \leftarrow X \rightarrow Y$,
2. $D \leftarrow F \rightarrow M \rightarrow Y$,

3. $D \leftarrow F \leftarrow U \rightarrow X \rightarrow Y$.

By conditioning on both F and X, we are said to *block* all such paths, allowing us to identify the average causal effect of D on Y, assuming no additional unobserved confounding.

Wrap-Up and Implications for 401(k) Analysis

Both functional and graphical perspectives yield the same conclusion. Functionally, once F and X are fixed, $Y(d)$ is governed solely by noise terms independent of those influencing the natural value of D. Graphically, conditioning on F and X blocks all paths from D to $Y(d)$ in the counterfactual DAG, and in the original DAG, all backdoor paths are blocked by $\{F, X\}$. In either case, we have

$$Y(d) \perp\!\!\!\perp D \mid F, X,$$

which formalizes the idea that, once F and X are taken into account, the naturally generated D is irrelevant for the counterfactual outcome $Y(d)$. This conditional ignorability allows us to identify the average causal effect of D on Y.

7.3 Definitions of General DAGs and ASEMs

The purpose of this section is to generalize the previous example to encompass general ASEMs and DAGs. Here, we provide concise general definitions and present key mathematical results.[10]

10: Although we believe the previous examples effectively illustrate the core concepts, presenting general definitions and results remains essential, given the foundational role of ASEMs and DAGs in causal inference.

A graph G is an ordered pair (V, E), where $V = \{1, \ldots, J\}$ is a set of vertices (nodes) and E is a collection of edges, represented by entries $e_{ij} \in \{0, 1\}$ for $(i, j) \in V \times V$.

Given a collection of random variables $X = (X_j)_{j \in V}$, we associate each index j with X_j and use them interchangeably for convenience. If an edge (i, j) exists (i.e., $e_{ij} = 1$), we interpret it as

$$\text{“}X_i \rightarrow X_j\text{”} \quad \text{or} \quad \text{“}X_i \text{ is an immediate cause of } X_j\text{”}.$$

Consider a strict partial order $<$ on V induced by E, where $X_j < X_k$ (read as “X_j is determined before X_k”) means either

$X_j \to X_k$ or there exists a directed path

$$X_j \to X_{v_1} \to \cdots \to X_{v_m} \to X_k.$$

A partial ordering exists if no node precedes itself (i.e., the graph contains no cycles).[11]

11: The absence of cycles ensures that $X_j < X_j$ never holds.

Definition 7.3.1 (DAG) *The graph* $\mathsf{G} = (V, E)$ *is a* directed acyclic graph (DAG) *if it contains no cycles; equivalently, if* V *is partially ordered by the edge structure* E*.*

Definition 7.3.2 (Parents, Ancestors, and Descendants) *The* parents *of* X_j *are defined as*

$$Pa_j := \{X_k : X_k \to X_j\}.$$

The children *of* X_j *are*

$$Ch_j := \{X_k : X_j \to X_k\}.$$

The ancestors *of* X_j *are*

$$An_j := \{X_k : X_k < X_j\} \cup \{X_j\},$$

and the descendants *of* X_j *are*

$$Ds_j := \{X_k : X_k > X_j\}.$$

From DAGs to ASEMs

Every causal DAG implicitly defines a nonparametric acyclic structural equation model (ASEM); the two are equivalent representations of the same assumptions about the data-generating process. In this perspective, DAGs serve as visual depictions of ASEMs, while ASEMs provide the structural equation formulations of DAGs.

Definition 7.3.3 (ASEM) *The ASEM corresponding to the DAG* $\mathsf{G} = (V, E)$ *is the collection of random variables* $\{X_j\}_{j \in V}$ *satisfying*

$$X_j := f_j(Pa_j, \epsilon_j), \quad j \in V,$$

where the disturbances $(\epsilon_j)_{j \in V}$ *are jointly independent.*

Definition 7.3.4 (Linear ASEM) *A* linear ASEM *is an ASEM in which the equations are linear:*

$$f_j(Pa_j, \epsilon_j) := f_j' Pa_j + \epsilon_j.$$

Here, the functions $\{f_j\}$ *are identified with their coefficient vectors.*

In linear ASEMs, the requirement of independent errors may be relaxed to uncorrelated errors.

Definition 7.3.5 (Structural /Potential Response Processes) *The* structural potential response processes *for the ASEM corresponding to* $\mathsf{G} = (V, E)$ *are given by*

$$X_j(pa_j) := f_j(pa_j, \epsilon_j), \quad j \in V,$$

viewed as stochastic processes indexed by the potential parental values pa_j.

Definition 7.3.6 (Consistency) *The observable variables are generated by drawing* $\{\epsilon_j\}_{j\in V}$ *and then solving the system of equations for* $\{X_j\}_{j\in V}$.

The stochastic shocks $\{\epsilon_j\}_{j\in V}$ are called *exogenous variables*, while the variables $\{X_j\}_{j\in V}$ are *endogenous*; the latter are determined by the model equations, whereas the former are not.

The joint distribution of variables in an ASEM is characterized by the following theorem.

Theorem 7.3.1 (Factual Law via Markovian Factorization) *The ASEM* $(X_j)_{j\in V}$ *associated with a DAG* $\mathsf{G} = (V, E)$ *satisfies the following equivalent properties:*

- ***Factorization:***

$$\mathsf{p}(\{x_\ell\}_{\ell\in V}) = \prod_{\ell\in V} \mathsf{p}(x_\ell \mid pa_\ell).$$

- ***Local Markov Property:*** *Each variable is independent of its non-descendants given its parents.*

D-Separation and Testable Restrictions

Next, we examine the constraints on the data-generating process implied by a given DAG. We introduce the concept of *d-separation*

and demonstrate that it implies conditional independence, known as the global Markov condition associated with the DAG. To proceed, we first require several definitions.

Definition 7.3.7 (Paths and Backdoor Paths on DAGs) *A* directed path *is a sequence*

$$X_{v_1} \to X_{v_2} \to \cdots \to X_{v_m}.$$

A non-directed path *is a path in which some, but not all, arrows are replaced by* $\leftarrow$*. A node* X_j *is a* collider *on a path if the path includes a subpath of the form* $\to X_j \leftarrow$*. A* backdoor path *from* X_l *to* X_k *is a non-directed path that starts at* X_l *and ends with an arrow into* X_k*.*

Definition 7.3.8 (Blocked Paths) *A path* π *is* blocked *by a set of nodes* S *if either:*

1. π *contains a chain* $i \to m \to j$ *or a fork* $i \leftarrow m \to j$ *with* $m \in S$*, or*
2. π *contains a collider* $i \to m \leftarrow j$ *such that neither* m *nor any descendant of* m *is in* S*.*

A path that is not blocked is open*.*

In Figure 7.3, the backdoor path $Y \leftarrow X \to D$ is blocked by setting $S = X$.

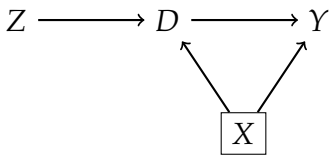


Figure 7.3: The path $Y \leftarrow X \to D$ is blocked by conditioning on X.

Definition 7.3.9 (Opening a Path by Conditioning) *A path containing a collider is* opened *by conditioning on that collider or one of its descendants.*

In Figure 7.4, the path $Y \to C \leftarrow D$ is blocked (by the empty set) but becomes open when conditioned on the collider C.

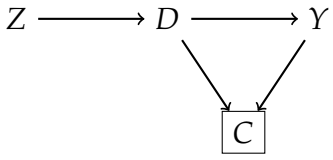


Figure 7.4: The path $Y \to C \leftarrow D$ is blocked but opens when conditioning on C.

Definition 7.3.10 (d-Separation) *Given a DAG* G*, a set of nodes* S d-separates *nodes* X *and* Y *if* S *blocks all paths between* X *and* Y*. We denote this as*

$$(X \perp\!\!\!\perp_d Y \mid S)_{\mathsf{G}}.$$

The following theorem establishes a fundamental link between d-separation and conditional independence.

Theorem 7.3.2 (Verma and Pearl [5]; Conditional Indepen-

dence from d-Separation) *If* $(X \perp\!\!\!\perp_d Y \mid S)_{\mathsf{G}}$ *holds, then*

$$X \perp\!\!\!\perp Y \mid S.$$

Intuitively, conditioning on S interrupts the information flow between X and Y, rendering them unable to predict each other given S. While this result is intuitive and verifiable in simple cases, its formal proof is nontrivial. The converse does not generally hold but is argued to hold "generically" (see Section 7.C).

Example 7.3.1 We illustrate how d-separation implies conditional independence:

1. In Figure 7.5, X and Y are d-separated by $S = \{Z, U\}$ since S blocks all paths between them. By Markov factorization,

$$\begin{aligned} \mathsf{p}(y, x \mid u, z) &= \mathsf{p}(y \mid x, z, u)\, \mathsf{p}(x \mid z, u) \\ &= \mathsf{p}(y \mid u, z)\, \mathsf{p}(x \mid z, u), \end{aligned}$$

   implying $X \perp\!\!\!\perp Y \mid Z, U$.
2. In Figure 7.6, X and Y are d-separated by $S = \{Z\}$, and similarly,

$$\mathsf{p}(y, x \mid z) = \mathsf{p}(y \mid z)\, \mathsf{p}(x \mid z),$$

   implying $X \perp\!\!\!\perp Y \mid Z$.

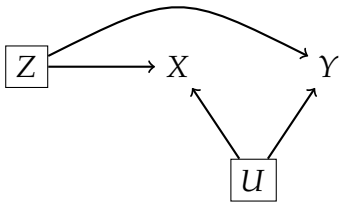


Figure 7.5: Example of d-separation.

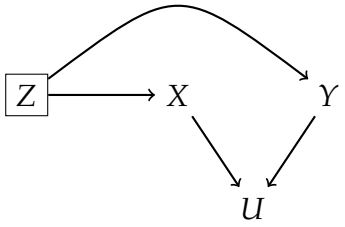


Figure 7.6: Example of d-separation.

These testable restrictions—known as *exclusion restrictions* in econometrics—can be expressed as

$$Y \perp\!\!\!\perp X \mid Z \quad \Longleftrightarrow \quad \mathsf{p}(y \mid x, z) = \mathsf{p}(y \mid z), \tag{7.3.1}$$

which is equivalent to

$$\mathrm{E}[g(Y) \mid X, Z] = \mathrm{E}[g(Y) \mid Z] \tag{7.3.2}$$

for any bounded function g. In other words, X does not improve the prediction of $g(Y)$ when Z is known. Numerous tests for such restrictions exist in the literature; see, e.g., [6].[12]

12: Examples include conditional independence tests, exclusion restriction tests, or conditional moment tests.

In linear ASEMs, these tests reduce to hypotheses about regression coefficients. For example, to test whether $Y \perp\!\!\!\perp X \mid Z$, one can examine whether the coefficient $\alpha = 0$ in the regression

$$Y = \alpha' X + \beta' Z + \epsilon, \quad \epsilon \perp Z.$$

Standard statistical tools can implement such tests; see the R Dagitty Notebook and the Python Pgmpy Notebook 7.7.1 for

examples.

Remark 7.3.1 (Equivalence of Local and Global Markov Properties) The local Markov property, Markov factorization, and global Markov property are equivalent [7]. Thus, any of these properties can be used to assess the validity of the Markov structure.

7.4 Counterfactuals and Identification by Conditioning

Counterfactuals

In this subsection, we focus on counterfactuals induced by *fix* interventions, a concept that builds on the foundation of causal reasoning.[13] This approach allows us to explore hypothetical scenarios while preserving the underlying structure of the original model, making it particularly useful for understanding causal effects in complex systems.

13: A fix-intervention extends the do-intervention by retaining the natural version of the variable while creating an intervention version. In contrast, the original do-intervention, introduced by Pearl, erases the natural version by replacing it entirely with the intervention version.

Definition 7.4.1 (Counterfactual ASEM Induced by a Fix Intervention) *The intervention* $\text{fix}(X_j = x_j)$ *on an ASEM creates a counterfactual ASEM (CF-ASEM) defined by a modified DAG, known as a Single World Intervention Graph (SWIG):*

$$\tilde{\mathsf{G}}(x_j) := (\tilde{V}, \tilde{E}),$$

along with a collection of counterfactual variables $\{X_k^*\}_{k \in V} \cup \{X_a^*\}$. *Here, the node* X_j *is split into two distinct entities:* $X_j^* := X_j$, *representing the natural value, and a new deterministic node a with* $X_a^* := x_j$, *representing the intervened value. The construction proceeds as follows:*

- *The intervention node* X_a^* *inherits only the outgoing edges from* X_j *(i.e.,* $\tilde{e}_{ai} = e_{ji}$ *for all i) and has no incoming edges (*$\tilde{e}_{ia} = 0$ *for all i), reflecting that it is fixed by the intervention.*
- *The node* X_j^* *inherits only the incoming edges from* X_j *(i.e.,* $\tilde{e}_{ij} = e_{ij}$ *for all i) and has no outgoing edges (*$\tilde{e}_{ji} = 0$ *for all i), preserving its dependence on its original causes.*
- *All remaining edges are preserved:* $\tilde{e}_{ik} = e_{ik}$ *for all* i *and for all* $k \neq j, k \neq a$, *ensuring the rest of the graph structure remains intact.*

► *The counterfactual variables are assigned according to*

$$X_k^* := f_k(Pa_k^*, \epsilon_k), \quad \text{for } k \neq a,$$

where Pa_k^* denotes the parents of X_k^* under $\tilde{E}$, adapting the structural equations to the new graph.

Interventions like $\text{fix}(X_j = x_j)$ induce new counterfactual distributions for the endogenous variables, offering a window into what would happen under specific conditions. Non-descendants of X_j remain unchanged, so $X_k^* = X_k$ for all X_k that are not downstream of X_j. For simplicity, we drop the "stars" from counterfactual variables that exactly replicate their factual counterparts, streamlining notation where the intervention has no effect.

Ignorability by D-Separation in Counterfactual DAGs

Consider any variable D in an ASEM as a treatment of interest and one of its descendants Y as an outcome we wish to study. Our goal is to identify the causal effect of D on Y, which requires finding an adjustment set S that ensures conditional exogeneity, or ignorability. This condition is formally expressed as:

$$Y(d) \perp\!\!\!\perp D \mid S,$$

meaning that, given S, the counterfactual outcome $Y(d)$ is independent of the natural value of D. This independence allows us to isolate the causal impact of setting $D = d$.

To achieve this, we construct the counterfactual DAG induced by the $\text{fix}(D = d)$ intervention, which replaces D with the fixed value d in all structural equations defining its descendants. This modified graph, or SWIG, reflects the system under the intervention. If, in this SWIG, $Y(d)$ is d-separated from D by a set S, then the conditional exogeneity condition holds:

$$Y(d) \perp\!\!\!\perp D \mid S.$$

D-separation here means that all paths between D and $Y(d)$ are blocked by S, ensuring no confounding influences remain.

Given conditional exogeneity, we can identify counterfactual expectations, such as $\mathrm{E}[g(Y(d)) \mid S = s]$, from observed data,

specifically by the regression $\mathrm{E}[g(Y) \mid S = s, D = d]$, provided the positivity condition $\mathsf{p}(s, d) > 0$ holds. This positivity ensures that every combination of $S = s$ and $D = d$ we condition on is observable in the data. By exogeneity and consistency, we have:

$$\mathrm{E}[g(Y(d)) \mid S = s] = \mathrm{E}[g(Y) \mid S = s, D = d],$$

and integrating over S gives the average potential outcome:

$$\mathrm{E}[g(Y(d))] = \mathrm{E}\Big[\mathrm{E}[g(Y) \mid S, D = d]\Big],$$

assuming $\mathsf{p}(s, d) > 0$ for all s in the support of $S \mid D = d$. This process links hypothetical outcomes to measurable quantities.

The following theorem, essentially due to [8], formalizes this approach.

Theorem 7.4.1 (A Counterfactual Criterion for Identification by Conditioning) *Consider any ASEM with DAG* G*. Re-label the treatment node* X_j *as* D*, and let* Y *be any descendant of* D *representing the outcome. Construct the SWIG* $\tilde{\mathsf{G}}(d)$ *induced by the* $\mathrm{fix}(D = d)$ *intervention, and let* S *be any subset of nodes common to both* G *and* $\tilde{\mathsf{G}}(d)$ *such that* $Y(d)$ *is d-separated from* D *by* S *in* $\tilde{\mathsf{G}}(d)$*. Then:*

- *The conditional exogeneity condition holds:*

$$Y(d) \perp\!\!\!\perp D \mid S.$$

- *The conditional average potential outcome is identified by the corresponding regression:*

$$\mathrm{E}[g(Y(d)) \mid S = s] = \mathrm{E}[g(Y) \mid D = d, S = s],$$

for all s *with* $\mathsf{p}(s, d) > 0$ *and for all bounded functions* g*.*

This criterion is "complete" in that it captures all valid adjustment sets excluding descendants of D. As discussed in [8], it encompasses all adjustment sets verifiable through an implementable intervention, providing a robust tool for causal inference in practice.

Example 7.4.1 (Pearl's Example) Consider the DAG in Figure 11.1 (introduced earlier as Pearl's Example) and its corresponding ASEM (not explicitly written). We aim to estimate the causal effect of D on Y, i.e., the mapping $d \mapsto Y(d)$. The

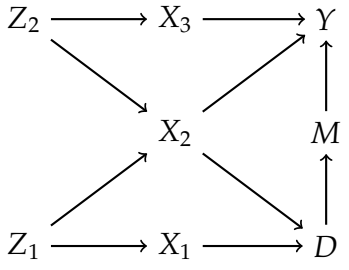


Figure 7.7: A DAG in Pearl's Example.

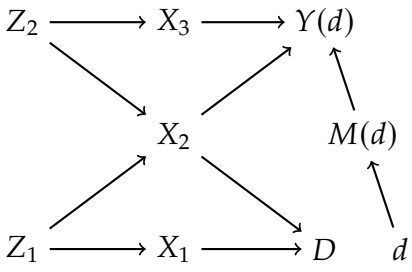


Figure 7.8: The DAG induced by the $\mathrm{fix}(D = d)$ intervention in Pearl's Example.

SWIG induced by the fix($D = d$) intervention is shown in Figure 7.8. In this counterfactual graph, valid adjustment sets S include:

$$\{X_1, X_2\}, \quad \{X_2, X_3\}, \quad \{X_2, Z_2\}, \quad \{X_2, Z_1\},$$

since each set blocks all open paths between $Y(d)$ and D. Conditioning on X_2 alone is insufficient because, although it blocks the inner backdoor paths, it opens an outer path where X_2 acts as a collider; adding X_1, X_3, Z_1, or Z_2 blocks this additional path, ensuring ignorability.

Remark 7.4.1 (Useful Limitation of the Counterfactual Criterion Approach*) A surprisingly useful limitation of the counterfactual DAG approach is that it avoids selecting valid yet unhelpful control variables for adjustment. Consider the simple DAG

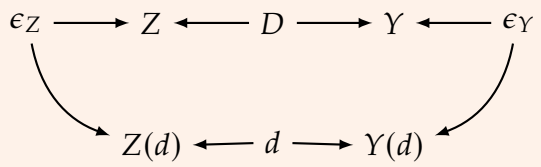
$$Z \leftarrow D \rightarrow Y,$$

and its corresponding counterfactual DAG

$$Z(d) \leftarrow d \rightarrow Y(d).$$

Under the counterfactual approach, no adjustment is required—in other words, the empty set is a valid adjustment set: $Y(d) \perp\!\!\!\perp D$. However, we know that Z is a valid control.

We can deduce its validity by considering a *cross-world* DAG that combines factual and counterfactual variables from the respective ASEMs:



In this cross-world DAG, $Y(d)$ is d-separated from D by Z, so that

$$Y(d) \perp\!\!\!\perp D \mid Z.$$

Thus, while Z is a valid control variable, it is arguably superfluous, as it does not add useful information about $Y(d)$. In fact, adjusting for Z can reduce the precision of our estimates of the average causal effect of D on Y.

Ignorability by Backdoor Blocking in Factual DAG

Pearl [7] developed a powerful and practical criterion for establishing conditional exogeneity/ignorability by analyzing the structure of the factual DAG, without needing to construct a counterfactual DAG. This method, known as the backdoor criterion, simplifies the process of identifying valid adjustment sets and is widely applied in causal inference studies, especially when working directly with observed data. We note that the proof provided by Pearl [2] itself relies on the concept of d-separation within a counterfactual DAG.[14] This approach underscores the deep connection between the backdoor criterion and counterfactual reasoning.

14: Specifically, Pearl constructs a modified DAG by removing all outgoing edges from D and then examines whether the remaining paths from D to Y are blocked, meaning they are d-separated. This modified DAG is essentially equivalent to the counterfactual DAG used in the fix-intervention framework.

Theorem 7.4.2 (Backdoor Criterion) *Consider any ASEM and its associated DAG. Re-label a treatment node X_j as D, and let Y, an outcome of interest, be any descendant of D. The adjustment set S is valid, meaning it implies conditional ignorability*

$$Y(d) \perp\!\!\!\perp D \mid S,$$

if the backdoor criterion is satisfied: No element of S is a descendant of D, and all backdoor paths from Y to D are blocked by S.

A backdoor path, recall, is one that starts at D and ends with an arrow pointing into D, representing confounding influences. The key insight is that blocking these backdoor paths eliminates confounding, leaving only the direct and indirect causal channels from D to Y. This approach is intuitive and efficient because it focuses on the factual DAG structure.

Example 7.4.2 (Pearl's Example Again, Using the Backdoor Criterion) The graph in Figure 11.1 has two backdoor paths from D to Y: the inner path

$$D \leftarrow X_2 \rightarrow Y,$$

and the more complex outer path

$$D \leftarrow X_1 \leftarrow Z_1 \rightarrow X_2 \leftarrow Z_2 \rightarrow X_3 \rightarrow Y.$$

Conditioning on X_2 alone does not suffice to identify the causal effect of D on Y. While it blocks the inner backdoor path, it opens the outer path by conditioning on X_2, a collider in that sequence. To resolve this, we must also condition on an additional variable like X_1, X_3, Z_1, or Z_2 to block the

outer path. Thus, valid adjustment sets include $S_1 = \{X_1, X_2\}$ or $S_2 = \{X_2, X_3\}$. Identifying other valid sets is left as an exercise. Note that conditioning on M is invalid since M is a descendant of D, representing an intermediate outcome that could bias the effect estimate.

Applying the backdoor criterion systematically can yield all minimal adjustment sets needed for identification; see [7]. However, it may not capture every possible valid set. Let's revisit the DAG: $Z \leftarrow D \rightarrow Y$. Here, conditioning on Z does not satisfy the backdoor criterion because Z is descendant of D, but it is a valid control variable. Here D directly causes Y, and there is no need to condition on Z. If we do condition Z, it does not confound the direct effect. However, conditioning on Z may lower the precision with which we estimate the effect of D on Y, thereby making Z potentially unhelpful, but not invalid control. This highlights the fact that this "limitation" of the backdoor approach is useful in disregarding controls that are not useful. As we had seen, the same comment applies to the counterfactual approach.

7.5 Notes

We adopt the framework pioneered by J. Pearl in his seminal *Biometrika* paper [2], with a few minor adaptations to enhance its applicability. First, we emphasize fix interventions over do interventions because fix interventions preserve the original "natural" variable alongside the intervened version, providing a seamless transition to counterfactual DAGs. This preservation enables us to systematically deduce conditional ignorability conditions, such as $Y(d) \perp\!\!\!\perp D \mid S$, directly from the graph structure. Fix interventions, as an extension of Pearl's earlier do-interventions, were formalized by Heckman and Pinto [9] and by Robins and Richardson [8]. Notably, elements of fix interventions also appear in Pearl's original work [2], particularly in his theoretical analysis and proofs, where he employs graph manipulations known as the do-calculus to explore counterfactual scenarios.

The connection between structural equation models (SEMs) and potential outcomes is encapsulated in what J. Pearl [7] calls the First Law of Causal Inference. This principle asserts that SEMs fully induce potential outcomes, highlighting that no information is lost by beginning causal analysis with SEMs rather than potential outcomes directly. In fact, starting with an SEM offers a distinct advantage: it allows us to mathematically

articulate contextual assumptions and derive conditional ignorability, rather than simply assuming it upfront as is common in potential outcomes frameworks. For instance, in empirical 401(k) analyses, researchers often claim that potential outcomes are independent of 401(k) eligibility given worker characteristics. However, constructing a comprehensive DAG that captures the full context reveals additional factors—such as firm characteristics—that must also be conditioned on to ensure ignorability. Thus, DAGs serve as a powerful tool to uncover critical details that might be overlooked in a less structured potential outcomes approach, grounding causal inference in a clearer and more explicit model of the data-generating process.

The uses of DAGs in empirical work are very common in epidemiology, see e.g. [10] for a review, common in theoretical work in computer science, and is much less common in empirical economics, despite the first use dating back to 1928 in the foundational work of Philip Wright [11]; there are recent attempts to revive the interest in economics, see [12] and [13].

7.6 Additional Resources

Dagitty.Net is an excellent online resource for plotting and analyzing causal DAG models. It contains many interesting examples used in empirical analyses across various fields.

Causalfusion.Net is another valuable online resource for exploring causal DAG models, covering several deviations from the standard framework.

7.7 Notebooks

Notebook 7.7.1 (DAGs I) R: Dagitty Notebook employs the R package "dagitty" to analyze Pearl's Example (Figure 11.1) as well as simpler ones. Python: Pgmpy Notebook employs the analogue with Python package "pgmpy" and conducts the same analysis. Both packages automatically list all conditional independence in a DAG; these are obtained by using the graphical d-separation criterion. We then go ahead and test those restrictions assuming a linear ASEM structure. The notebook also illustrates the analysis from the next chapter.

Notebook 7.7.2 (DAGs II) R: Dosearch Notebook employs the R-package "dosearch" to analyze Pearl's Example (Figure

11.1). This package automatically finds identification answers to causal queries, allowing us to also answer these types of queries under different data sources, sample selection, and other deviations from the standard framework. Python: Dosearch Notebook does the same thing by loading the R "dosearch" package into Python.

7.8 Exercises

Exercise 7.8.1 (401(k) Example) Consider the DAG for the 401(k) example, but suppose now that there is an additional arrow from U to D.

1. Write down the ASEM corresponding to the DAG. State the joint distribution of all variables in the model, exploiting the Markovian structure. What conditional independence restrictions are implied by the model? Are they testable?
2. Provide the counterfactual DAG (SWIG) for this DAG corresponding to the Fix intervention and state the corresponding ASEM. Using d-separation, determine sufficient adjustment sets for identifying the average causal effect of D on Y.
3. In the factual DAG, list all backdoor paths from D to Y. Identify sufficient adjustment sets for the average causal effect of D on Y using the "blocking backdoor paths" approach.
4. Now suppose there is no arrow from F to M, meaning the match amount does not statistically depend on firm characteristics. Determine the sufficient adjustment sets using either the counterfactual approach or the "blocking backdoor paths" approach. Which approach do you find easier to use?

Exercise 7.8.2 (Pearl's Example) Consider the DAG in Figure 11.1. Answer the following questions. The best way to answer these question is to use computational packages (but please explain the principles the package is using).

1. What are the testable implications of the assumptions embedded in the model? Hint: The testable implications are derived from the d-separation criterion.
2. Assume that only variables D, Y, X_2 and M are measured, are there any testable implications?
3. Now assume only D, Y, and X_2 are measured. Are

there any testable implications?
4. Now assume that all of the variables but X_2 (7 in total) are measured. Are there any testable restrictions?
5. Assume that an alternative model, competing with Model 1, has the same structure, but with the $X_2 \to D$ arrow reversed. What statistical test would distinguish between the two models?

Exercise 7.8.3 Work through the proof that d-separation implies conditional independence in Section 7.B. Supply the steps of the proof that were left as a homework or reading exercise.

7.A Review of Conditional Independence

The following lemma reviews various ways in which conditional independence can be established.

Lemma 7.A.1 (Equivalent Forms of Conditional Independence) Variables X and Y are conditionally independent given Z if and only if one of the following conditions is met:

1. $\mathrm{p}(x \mid y, z) = \mathrm{p}(x \mid z)$ if $\mathrm{p}(y, z) > 0$.
2. $\mathrm{p}(x \mid y, z) = f(x, z)$ for some function f.
3. $\mathrm{p}(x, y \mid z) = \mathrm{p}(x \mid z)\mathrm{p}(y \mid z)$ if $\mathrm{p}(z) > 0$.
4. $\mathrm{p}(x, y \mid z) = f(x, z)g(y, z)$ for some functions f and g.
5. $\mathrm{p}(x, y, z) = \mathrm{p}(x \mid z)\mathrm{p}(y \mid z)\mathrm{p}(z)$ if $\mathrm{p}(z) > 0$.
6. $\mathrm{p}(x, y, z) = \mathrm{p}(x, z)\mathrm{p}(y, z)/\mathrm{p}(z)$ if $\mathrm{p}(z) > 0$.
7. $\mathrm{p}(x, y, z) = f(x, z)g(y, z)$ for some functions f and g.

As a reading exercise prove the equivalence of (1) and (2), of (1) and (7), and of any other pair.

7.B Theoretical Details of d-Separation*

Here we explain why d-separation implies conditional independence.[15]

Lemma 7.B.1 (Easy Form of d-Separation) Let X, Y, and Z be three disjoint sets of variables in an ASEM such that their union is an ancestral set, that is, for any $X \in \mathsf{X} \cup \mathsf{Y} \cup \mathsf{Z}$ and

15: We follow the proof sketch presented in Nevin L. Zhang's lecture notes, but rely on ASEMs to simplify some arguments and supply a proof for a key claim.

$X' < X$ we have $X' \in \mathsf{X} \cup \mathsf{Y} \cup \mathsf{Z}$. If Z d-separates X and Y, then

$$\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}.$$

Proof. Let Z_1 be the set of nodes in Z that have parents in X. And let $\mathsf{Z}_2 = \mathsf{Z} \backslash \mathsf{Z}_1$.

Because Z d-separates X and Y, we have that (see Figure 7.9):

- For any $W \in \mathsf{X} \cup \mathsf{Z}_1$, $Pa_W \subseteq \mathsf{X} \cup \mathsf{Z}$;[16]
- For any $W \in \mathsf{Y} \cup \mathsf{Z}_2$, $Pa_W \subseteq \mathsf{Y} \cup \mathsf{Z}$.[17]

Let U denote the set of variables not included in X, Y, or Z. We then obtain a factorization

$$\begin{aligned} \mathrm{p}(\mathsf{x}, \mathsf{z}, \mathsf{y}) &= \int \prod_{W \in \mathsf{U} \cup \mathsf{X} \cup \mathsf{Y} \cup \mathsf{Z}} \mathrm{p}(w \mid Pa_W = pa_W) d\mathsf{u} \\ &= \int \prod_{W \in \mathsf{U}} \mathrm{p}(w \mid Pa_W = pa_W) d\mathsf{u} \\ &\quad \times \prod_{W \in \mathsf{X} \cup \mathsf{Z}_1} \mathrm{p}(w \mid Pa_W = pa_W) \\ &\quad \times \prod_{W \in \mathsf{Z}_2 \cup \mathsf{Y}} \mathrm{p}(w \mid Pa_W = pa_W), \end{aligned}$$

where in the last equality we used the fact that u does not appear at all in the second and third factors, since $\mathsf{X} \cup \mathsf{Y} \cup \mathsf{Z}$ is ancestral. Moreover, the second factor is a function of x and z alone and the third factor is a function of y and z alone. The integral is 1 by total probability.[18] It follows that $\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}$.[19] □

Now we restate the main claim we'd like to demonstrate, which is that d-separation implies conditional independence.

Global Markov. Let X and Y be two variables and Z be a set of variables that does not contain X or Y. If Z d-separates X and Y, then

$$X \perp\!\!\!\perp Y \mid \mathsf{Z}$$

Proof of Theorem 7.3.2.

Let X be the set of all ancestors of $\{X, Y\} \cup \mathsf{Z}$ that are *not* d-separated from X by Z. Let Y be the set of all ancestors of $\{X, Y\} \cup \mathsf{Z}$ that are neither in X nor in Z.

Key Claim: The set Z d-separates the sets X and Y.

The claim follows from the careful use of the definition of d-separation, and is proven below.

16: Suppose that any such node has a parent in Y. If it were a node in X, then we get a violation of d-separation. If it were a node in Z_1, then we have that Z_1 has one parent in X and one parent in Y and therefore it is a collider that was included in Z, violating d-separation.

17: Suppose that any such node has a parent in X. By the definition of Z_1 it has to be a node in Y. But then we have that a node in Y has a parent in X, violating d-separation.

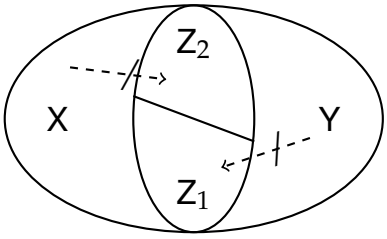


Figure 7.9: Pictorial representation of key argument in Lemma 7.B.1.

18: Prove this as a reading exercise by integrating over the variables in U in reverse order with respect to the DAG ordering.

19: Prove this as a reading exercise, i.e., prove bullet (7) of Lemma 7.A.1.

Given the key claim, Lemma 7.B.1 implies that $\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}$, since $\mathsf{X} \cup \mathsf{Y} \cup \mathsf{Z}$ is ancestral by its exhaustive construction. This implies that there must exist functions $f(\mathsf{x}, \mathsf{z})$ and $g(\mathsf{z}, \mathsf{y})$ such that

$$\mathrm{p}(\mathsf{x}, \mathsf{z}, \mathsf{y}) = f(\mathsf{x}, \mathsf{z}) g(\mathsf{z}, \mathsf{y}).$$

Since X is in X and Y in Y, the conclusion is reached.[20] □

20: Prove this explicitly, as a reading exercise, by integrating over all variables in $\mathsf{X} \backslash \{X\}$ and $\mathsf{Y} \backslash \{Y\}$ and invoking Lemma 7.A.1.

Proof of the Key Claim. Suppose that Z does not d-separate the sets X and Y and that there exists a node $X' \in \mathsf{X}$ which is not d-separated from some node $Y' \in \mathsf{Y}$. Thus, there is an open path X - - X',[21] and an open path X' - - Y'. Consider the concatenation of these two paths. If X' is not a collider on this concatenated path, then the path X - - X' - - Y' is also open, and therefore X is not d-separated from Y', which is in contradiction with the definition of X and Y. Thus X' has to be a collider on this concatenated path. Moreover, note that since we are only restricting our analysis to the ancestral set $An_{\{X,Y\} \cup \mathsf{Z}}$, we have that X' must be an ancestor of either Z or Y or X:

21: In this proof, we denote with U - - V a path from a node U to a node V and with $U \dashrightarrow V$ a directed path from U to V.

If X' is an ancestor of some node in Z then the path X - - X' - - Y' is again open, leading to a contradiction with the definition of X and Y.

If X' is an ancestor of Y, then there is a directed path $X' \dashrightarrow Y$. If that path is open, then there is an open path X - - $X' \dashrightarrow Y$, violating the fact that Z was d-separating X from Y. For the path to be closed, it must be that some node $Z \in \mathsf{Z}$ is on the path. However, in this case X' is an ancestor of a node in Z, which has already been excluded.

Finally, if X' is an ancestor of X, then there exists a directed path $X' \dashrightarrow X$. This path also has to be open, as if a node in Z existed on that path, then X' would be an ancestor of a node in Z, which has been excluded. However, in this case, we have an open path Y' - - $X' \dashrightarrow X$, from Y' to X, which violates the definition of X and Y. □

7.C Faithfulness and Causal Discovery

Given that DAGs effectively encode conditional independence relations, it is tempting to try to infer conditional independence directly from the data. *Causal discovery* refers to methods that indeed attempt to learn conditional independence relationships from data with one application being attempting to recover causal structures. The possibility of recovering causal structures

perfectly from the population data critically relies on the concept of faithfulness.

Recall that d-separation implies conditional independence, but the reverse implication

$$Y \perp\!\!\!\perp X \mid S \implies (Y \perp\!\!\!\perp_d X \mid S)_{\mathsf{G}} \tag{7.C.1}$$

is not true in general. If we restrict attention to the set of distributions p of random variables associated with graph G such that implication (7.C.1) holds, we are said to impose the *faithfulness* assumption on p.

Example 7.C.1 (Unfaithfulness) A trivial example is the DAG

$$X \to Y$$

where

$$Y := \alpha X + \epsilon_Y; \quad X := \epsilon_X;$$

with ϵ_X and ϵ_Y independent standard normal variables. Consider S to be the empty set. In this model we have that $Y \perp\!\!\!\perp X$ when $\alpha = 0$, but Y and X are not d-separated in the DAG $X \to Y$. The distribution p of (Y, X) corresponding to $\alpha = 0$ is said to be unfaithful. However, the exceptional point $\alpha = 0$ has a measure 0 on the real line, so this exception is said to be non-generic.

The observation about the simple example above generalizes: If probabilities p themselves are viewed as generated by Nature as a draw from a continuum P, where each $\mathsf{p} \in \mathsf{P}$ factorizes according to G, then the set of models where the reverse implication (7.C.1) does not hold has measure zero. This observation motivates the argument that the faithfulness assumption is a weak requirement; that is, a given p is "very unlikely" to be unfaithful.

Remark 7.C.1 (Causal Discovery) The use of the faithfulness assumption should allow us to discover the equivalence class of the true DAG from the population distribution p: We can compute all valid conditional independence relations and then discover the equivalence class of DAGs. See, for example, the PC algorithm [14] for an explicit causal discovery algorithm and the review provided in [15]. We can then apply contextual knowledge to further orient the edges of the graph.

Even though the set of unfaithful distributions has measure zero, the neighborhood of this set may not be small in high-

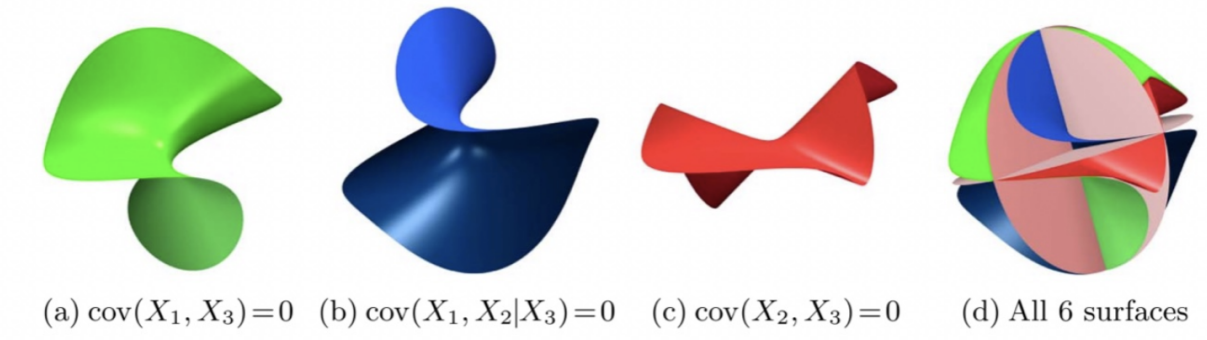


Figure 7.10: Uhler et. al [16]: A set of "unfaithful" distributions p in the simple triangular Gaussian SEM/DAG: $X_1 \to X_2, (X_1, X_2) \to X_3$.

dimensional graphs, which creates difficulty in inferring the DAG structure from an estimated version $\hat{\text{p}}$.

Example 7.C.2 (Unfaithfulness Continued) In the trivial example above, suppose that we have that $\hat{\alpha} = .1$ and $\hat{\alpha} \sim N(\alpha, \sigma^2)$ where $\sigma = .1$. Then we can't be sure whether $\alpha = 0$, $\alpha = .1$, or α equals any other number, though say a 95% confidence interval would have α between $-.1$ and .3. Therefore, we can't be sure whether the true model is

$$X \to Y \text{ or } X \quad Y.$$

Informally speaking, it is impossible to discover the true graph structure in this example when $\alpha \approx 0$. In econometrics jargon, this statement amounts to saying that we can't distinguish exact exclusion restrictions from "approximate" exclusion restrictions.

Thus, it is hard to distinguish exact independence from approximate independence with finite data. In high-dimensional graphs, the possibility that $\hat{p}$ lands in the "near-unfaithful" regions can be substantial, as Uhler et. al.[16]'s analysis shows.[22]

The observations above motivate a form of sensitivity analysis – e.g., Conley et al. [17] – where one replaces exact exclusion restrictions by approximate exclusion restrictions that can't be distinguished from exact exclusion restrictions and examines the sensitivity of causal effect estimates.

22: See Uhler et al's [16] figure; reproduced in Figure 7.10. The set is parameterized in terms of the covariance of (X_1, X_2, X_3). The right panel shows the set of unfaithful distributions, and the three other panels show 3 of 6 components of the set. Each of the cases corresponds to the non-generic case which would make faithfulness fail, leading to discovery of the wrong DAG structure. While the exact setting where faithfulness would fail is non-generic, there are many distributions that are "close" to these unfaithful distributions. This observation means that, in finite samples, we are not able distinguish models that are close to the set of unfaithful distributions from unfaithful distributions and may thus also discover the wrong DAG structure and correspondingly draw incorrect causal conclusions.

Bibliography

[1] Judea Pearl and Dana Mackenzie. *The Book of Why*. Penguin Books, 2019 (cited on page 164).

[2] Judea Pearl. 'Causal diagrams for empirical research'. In: *Biometrika* 82.4 (1995), pp. 669–688 (cited on pages 165, 171, 181, 182).

[3] Trygve Haavelmo. 'The probability approach in econometrics'. In: *Econometrica* 12 (1944), pp. iii–vi+1–115 (cited on page 165).

[4] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. 'Long Story Short: Omitted Variable Bias in Causal Machine Learning'. In: *arXiv preprint arXiv:2112.13398* (2023) (cited on page 171).

[5] Thomas Verma and Judea Pearl. *Influence diagrams and d-separation*. Tech. rep. Cognitive Systems Laboratory, Computer Science Department, UCLA, 1988 (cited on page 175).

[6] Rajen D. Shah and Jonas Peters. 'The hardness of conditional independence testing and the generalised covariance measure'. In: *Annals of Statistics* 48.3 (2020), pp. 1514–1538 (cited on page 176).

[7] Judea Pearl. *Causality*. Cambridge University Press, 2009 (cited on pages 177, 181, 182).

[8] Thomas S. Richardson and James M. Robins. *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality*. Working Paper No. 128, Center for the Statistics and the Social Sciences, University of Washington. 2013. URL: https://csss.uw.edu/files/working-papers/2013/wp128.pdf (cited on pages 179, 182).

[9] James Heckman and Rodrigo Pinto. 'Causal analysis after Haavelmo'. In: *Econometric Theory* 31.1 (2015 (NBER 2013)), pp. 115–151 (cited on page 182).

[10] Peter WG Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, et al. 'Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations'. In: *International journal of epidemiology* 50.2 (2021), pp. 620–632 (cited on page 183).

[11] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. New York: The Macmillan company, 1928 (cited on page 183).

[12] Paul Hünermund and Elias Bareinboim. 'Causal inference and data fusion in econometrics'. In: *The Econometrics Journal* (2023), utad008 (cited on page 183).

[13] Jaap H Abbring, Victor Chernozhukov, and Iván Fernández-Val. 'Philip G. Wright, directed acyclic graphs, and instrumental variables'. In: *Econometrics Journal* (2025) (cited on page 183).

[14] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000 (cited on page 188).

[15] Clark Glymour, Kun Zhang, and Peter Spirtes. 'Review of causal discovery methods based on graphical models'. In: *Frontiers in Genetics* 10 (2019), p. 524 (cited on page 188).

[16] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. 'Geometry of the faithfulness assumption in causal inference'. In: *Annals of Statistics* 41.2 (2013), pp. 436–463 (cited on page 189).

[17] Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi. 'Plausibly exogenous'. In: *Review of Economics and Statistics* 94.1 (2012), pp. 260–272 (cited on page 189).