Applied Causal Inference Powered by ML and AI

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

July 28, 2024

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 associated structural equation models. These provide a useful tool for thinking about structures underlying causal identification.

7.1 Introduction

The purpose of this module is to provide a more formal and general treatment of acyclic nonlinear (and nonparametric) structural equation models (SEMs) and corresponding causal directed acyclic graphs (DAGs). We discuss the concepts and identification results provided by Judea Pearl and his collaborators and by James M. Robins and his collaborators.

In 2011, J. Pearl was awarded the A.M. Turing award, the highest award in the field of Computer Science and Artificial Intelligence: "For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning." In the Biometrika 1995 article [2], J. Pearl presents his work as a generalization of the SEMs put forward by T.Haavelmo [3] in 1944 and others.

These models and concepts allow us to rigorously define structural causal effects in fully nonlinear models and obtain conditional independence relationships that can be used as inputs to establishing nonparametric identification from the structure of the causal DAGs alone.[1] Structural causal effects are defined as hypothetical effects of interventions in systems of equations. We discuss identification of effects of *do interventions* introduced by Pearl [2] and *fix interventions* introduced by Heckman and Pinto [4] and Robins and Richardson [5].[2] fix interventions induce counterfactual DAGs called SWIGs (Single World Intervention Graphs) and can recover the causal graphs we've seen in previous chapters.

1: We abstract away from rank-type conditions. See Remark 6.2.1.

2: Fix interventions also had appeared as part of do calculus in Pearl [2].

Whether causal effects derived from SEMs approximate policy or treatment effects in the real world depends to a large extent on the degree to which the posited SEM approximates real phenomena. In thinking about the approximation quality of a model, it is important to keep in mind that we will never be able to establish that a model is fully correct using statistical criteria. However, we may be able to reject a given model using formal falsifiability criteria – though not all models are statistically falsifiable – or contextual knowledge. Further, evidence for some causal effects inferred from SEMs can be provided by further use of explicit randomized controlled trials, though the use of experiments is not an option in many cases. Ultimately, contextual knowledge is often crucial for making the case that a given structural model represents real phenomena sufficiently well to produce credible estimates of causal effects when using observational data.

Notation

Consider a pair of random variables (or equivalently, random vectors) U and V with joint distribution probability (mass) function $\mathsf{p}_{UV}(u, v)$ at generic evaluation points (u, v). We will simply denote $\mathsf{p}_{UV}(u, v)$ by $\mathsf{p}(u, v)$ whenever there is no ambiguity. We will denote the marginal probability (mass) functions

by $\mathsf{p}_U(u)$ and $\mathsf{p}_V(v)$, or simply by $\mathsf{p}(u)$ and $\mathsf{p}(v)$. The random variables U and V are independent, which we denote as

$$U \perp\!\!\!\perp V,$$

if and only if the joint probability density (or mass) function $\mathsf{p}(u, v)$ can be factorized as

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

or equivalently if and only 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:

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

Conditional independence is defined similarly by replacing distributions and expectations with their conditional analogs. Appendix 7.B reviews some useful results on conditional independence.

7.2 From Causal Diagrams to Causal DAGs: TSEM Example

Formal causal nonlinear DAGs generalize linear parametric models to general nonparametric forms. Recall our previous discussion of a model for a household's log-demand for gasoline (Y), which is a function of log-price (p) and household characteristics (X). We can generalize the simple TSEM to a nonlinear DAG as follows.

Example 7.2.1 (TSEM) We have a system of triangular structural equations:

$$\begin{aligned} Y &:= f_Y(P, X, \epsilon_Y), \\ P &:= f_P(X, \epsilon_P), \\ X &:= \epsilon_X, \end{aligned} \tag{7.2.1}$$

where f's are said to be deterministic structural functions and $\epsilon_Y, \epsilon_P, \epsilon_X$ are structural shocks that are independent of each other. The dimension of structural shocks is not restricted. Also, note the independences:

$$\epsilon_Y \perp\!\!\!\perp (P, X), \quad \epsilon_P \perp\!\!\!\perp X.$$

A causal diagram depicting the algebraic relationship defining the TSEM in Example 7.2.1 is shown in Figure 7.1. The absence of edges between nodes encodes the model's independence restrictions. Thus, as before, we can see that we can view graphs as representations of independence relations in statistical models. The graph visually depicts independence restrictions and the propagation of information or structural shocks from root nodes to their children, grandchildren, and so forth.

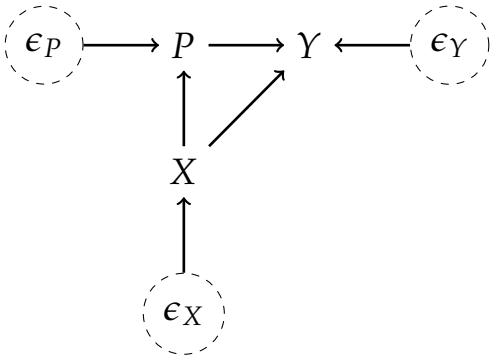

Figure 7.1: The causal DAG equivalent to the TSEM in Example 7.2.1.

It is also common to draw graphs based on only observed variables. We can erase the latent root nodes from Figure 7.1 to produce the equivalent diagram illustrated in Figure 7.2.

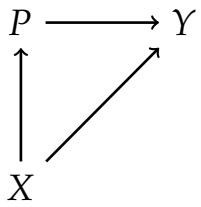

Figure 7.2: The causal DAG corresponding to the TSEM in Example 7.2.1 with latent root nodes erased.

The TSEM is purely a statistical model. We can view this model as structural under invariance restriction, following Haavelmo [3].

Definition 7.2.1 (Structural Form) *When we say that the TSEM is structural, we mean that it is defined by a structure made up of a set of stochastic processes:*

$$\begin{aligned} Y(p, x) &:= f_Y(p, x, \epsilon_Y), \\ P(x) &:= f_P(x, \epsilon_P), \\ X &:= \epsilon_X, \end{aligned}$$

indexed by $(p, x) \in \mathcal{P} \times \mathcal{X}$, called structural functions or structural potential outcome processes. Moreover,

- *(Exogeneity) Stochastic shocks ϵ_P, ϵ_X, and ϵ_Y are generated as independent variables outside of the model;*
- *(Consistency) The endogenous variables are generated by recursive substituations:*

$$Y := Y(P, X), \quad P := P(X), \quad X := \epsilon_X;$$

- *(Invariance) The structure remains invariant to changes of the distribution of stochastic shocks ϵ.*

The structure will be assumed to be preserved under various

interventions as defined below.

While SEMs are statistical models, assumptions akin to those in Definition 7.2.1 endow them with a structural meaning. Structural meaning may be generated by economic or other scientific reasoning. For example, structural functions may correspond to demand functions, supply functions, and expenditure functions, with these notions going back at least to Marshall [6] in the 19th century.

Remark 7.2.1 (Link to Potential Outcomes) Consider binary $p \in \{0, 1\}$ for simplicity. Consider potential outcomes, given by the structure:

$$Y(p, X) := g(p, X, \epsilon_Y(p)).$$

We can view potential outcomes through a SEM framework as follows. Let $\epsilon_Y := \{\epsilon_Y(p) : p \in \{0, 1\}\}$, then we have that

$$Y(p, X) = g(p, X, \epsilon_Y(p)) = f_Y(p, X, \epsilon_Y),$$

for

$$f_Y(p, x, e) := 1(p = 0)g(p, x, e(0)) + 1(p = 1)g(p, x, e(1))$$

for the argument $e = \{e(p) : p \in \{0, 1\}\}$. This example emphasizes that the dimensionality of ϵ's is not restricted in the general framework.

Identification by Regression

By conditioning on $X = x$ in the graph in Figure 7.1, we obtain the graph shown in Figure 7.3. We can equivalently express the relationship shown in Figure 7.3 in terms of equations as

$$Y(x) = f_Y(P(x), x, \epsilon_Y), \quad \epsilon_Y \perp\!\!\!\perp P(x).$$

If $P(x)$ is non-degenerate, we can further condition on $P(x) = p$ to learn the average structural function

$$\mathrm{E}[f_Y(p, x, \epsilon_Y)]$$

via regressions. We formally record this result as follows.

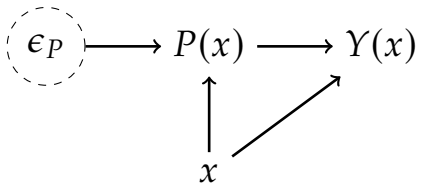

Figure 7.3: The graph produced from Figure 7.1 by conditioning on $X = x$. Here X is a parent to both P and Y. After conditioning, the remaining source of variation in $P(x)$ is ϵ_P. ϵ_P is determined exogenously – as if by an experiment – which allows measurement of the causal effect $P(x) \to Y$.

In the TSEM, the conditional average structural function

$$\mathrm{E}[f_Y(p, x, \epsilon_Y)]$$

can be identified by conditioning on P and X:

$$\begin{aligned}&\mathrm{E}[Y|P = p, X = x]\\ &\quad = \mathrm{E}[f_Y(P, X, \epsilon_Y)|P = p, X = x]\\ &\quad = \mathrm{E}[f_Y(p, x, \epsilon_Y)|P = p, X = x]\\ &\quad = \mathrm{E}[f_Y(p, x, \epsilon_Y)]\end{aligned}$$

provided the event $\{P = p, X = x\}$ is assigned positive density.

This average structural function has the interpretation as the expected outcome when P and X are exogenously set (set outside of the model as if by a policy maker or experiment) to $P = p$ and $X = x$.

Hence, we can use the average structural function to provide counterfactual predictions – predictions for the outcome under exogenous assignment of the policy variable P at fixed values for X. Within the TSEM, these counterfactual predictions align with the usual prediction rule $\mathrm{E}[Y|P = p, X = x]$.

If the confounder X is not observed, the causal relationship $P(x) \rightarrow Y$ is not identified.

If we can identify the conditional average structural function, we can also identify the conditional average structural causal effect:

$$\begin{aligned}&\mathrm{E}[f_Y(p_1,x, \epsilon_Y)] - \mathrm{E}[f_Y(p_0, x, \epsilon_Y)]\\ &\qquad = \mathrm{E}[Y|P = p_1, X = x] - \mathrm{E}[Y|P = p_0, X = x].\end{aligned} \tag{7.2.2}$$

The right hand side of (7.2.2) is a statistical quantity that can clearly be learned from data on Y, P, and X under reasonable assumptions. The left hand side of (7.2.2) defines a structural quantity of interest: the average effect of exogenously changing P from p_0 to p_1 at $X = x$.

Interventions

Do Interventions. The do operation $do(P = p)$ or do intervention corresponds to creating the counterfactual graph shown in Figure 7.4. On the graph, we remove P and replace it with a deterministic node p. In terms of equations (7.2.1) defining the TSEM, we replace the equation for P with p and then set P equal to p in the first equation. The corresponding counterfactual SEM is

$$\left(\begin{bmatrix} Y \\ P \\ X \end{bmatrix} : \text{do}(P = p)\right) := \begin{bmatrix} f_Y(p, X, \epsilon_Y) \\ p \\ X \end{bmatrix} = \begin{bmatrix} Y(p) \\ p \\ X \end{bmatrix}.$$

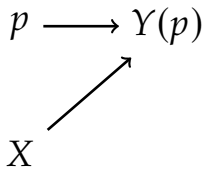

Figure 7.4: Causal DAG describing the counterfactual SEM induced by doing $P = p$.

The variables $Y(p)$ and X are the counterfactuals generated by the intervention $do(P = p)$. Note that the intervention keeps X and stochastic shocks ϵ_Y invariant.

The do operation has been extended to generate other types counterfactuals. For instance, another class of interventions are soft interventions[3] where the intervening variable is set to a value that is a function of its natural value (e.g., increasing a price by 10%). We could represent such interventions by the modified counterfactual SEM:

$$\left(\begin{bmatrix} Y \\ P \\ X \end{bmatrix} : \text{soft}_Y(P, \alpha)\right) := \begin{bmatrix} f_Y(\alpha(P), X, \epsilon_Y) \\ f_P(X, \epsilon_P) \\ X \end{bmatrix} = \begin{bmatrix} Y(\alpha(P)) \\ P \\ X \end{bmatrix}.$$

As an additional general example, we now consider *fix interventions* that induce single-world intervention graphs (SWIGs).[4]

Fix Interventions and SWIGs. Instead of removing P from the graph in Figure 7.2, we can split it into two nodes – P and a deterministic node p – where all the outgoing arrows from P are removed. The fixed node p then inherits the outgoing arrows from the original P.

The corresponding counterfactual SEM is

$$\left(\begin{bmatrix} Y \\ P \\ X \end{bmatrix} : \text{fix}_Y(P = p)\right) := \begin{bmatrix} f_Y(p, X, \epsilon_Y) \\ P \\ X \end{bmatrix} = \begin{bmatrix} Y(p) \\ P \\ X \end{bmatrix}.$$

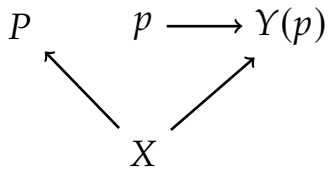

Figure 7.5: Causal DAG describing the counterfactual SEM induced by setting $P = p$ in the Y equation in (7.2.1) (formally a SWIG).

3: The ideas of constructing counterfactuals go back at least to P. Wright's work in 1928 [7], which involved replacing one structural equation with a different equation to define a counterfactual SEM. Specifically, Wright replaced the supply equation with another one reflecting a multiplicative tariff on the price that producers receive. This intervention is a (multiplicative) soft intervention. Building on P. Wright's work, soft interventions have been widely used in empirical economics (e.g., decomposition analysis of wages to study discrimination, carbon and emission taxes in environmental economics and industrial organization). See also [8, 9] for recent theoretical research in the computer science literature, framed in terms of DAGs and nonlinear ASEMs.

4: The fix intervention was introduced in Heckman and Pinto [4], as an extension of the do operation, and SWIGs were developed by Richardson and Robins [5].

The fix intervention merely says that we are setting $P = p$ in the Y equation. Figuratively speaking, it is a "localized do" operation. The variables $Y(p), P$, and X are the counterfactuals generated by this intervention. The intervention does not affect the P and X equations, nor does it affect ϵ_Y in the Y equation.

The SWIG allows us to immediately see that conditional exogeneity (ignorability) holds:

$$Y(p) \perp\!\!\!\perp P \mid X.$$

Therefore we can identify the counterfactual regression $\mathrm{E}[Y(p) \mid X]$ by the "factual" regression $\mathrm{E}[Y \mid P = p, X]$,

$$\mathrm{E}[Y(p) \mid X] = \mathrm{E}[Y(p) \mid P = p, X] = \mathrm{E}[Y \mid P = p, X],$$

invoking conditional independence and consistency arguments.

The do and fix interventions generate the same counterfactual distribution for $(Y(p), X)$, so the average causal effects of simple interventions coincide in the two approaches. However, the fix intervention creates a triple $(Y(p), X, P)$, which is useful for answering more complicated counterfactual questions.

For example, the counterfactual prediction $\mathrm{E}[Y(0) \mid P = 1]$ tells us what trainees ($P = 1$) would have earned on average, had they not gone through the training program ($p = 0$). In treatment effect analysis, this quantity is crucial for defining the average treatment effects for the treated:

$$\mathrm{E}[Y(1) \mid P = 1] - \mathrm{E}[Y(0) \mid P = 1].$$

Thus, the fix intervention allows us to seamlessly talk about conditional on P counterfactuals:[5]

$$\mathrm{E}\Big[Y(p) \mid P = \bar{p}\Big] := \mathrm{E}\Big[(Y \mid P = \bar{p}) : \mathrm{fix}_Y(P = p)\Big].$$

5: The same statement is formally not true with the do operation in place of the fix operation. Of course, one can also define these conditional counterfactuals by reverting to potential outcomes notation within causal DAGs; see [10].

7.3 General Acyclic SEMs and Causal DAGs

We will now turn to generalizing the concepts of the previous section from the TSEM case to general Directed Acyclic Graphs (DAGs) and the corresponding acyclic structural equation models (ASEMs).

DAGs and Acyclic SEMs via Examples

We now give a sequence of formal definitions, which can be easily understood by looking at just a single example.

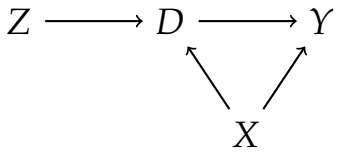

Figure 7.6: LS-DAG Example

Example 7.3.1 (Less Simple DAG (LS-DAG)) A directed acyclic graph (DAG) is a collection of nodes and directed edges with no cycles.

Consider the DAG in Figure 7.6: Here we can *say* that

- X is a parent of its children D and Y;
- D and Y are descendants of Z;
- There is a directed path from Z to Y;
- There are two paths from Z to X, but no directed path;
- D is a collider of the path $Z \to D \leftarrow X$;
- D is a noncollider of the path $Z \to D \to Y$;
- $Y \leftarrow X \to D$ is a backdoor path from Y to D.
- There are no cycles (there is no directed path that returns to the same node).

Example 7.3.2 (ASEM Corresponding to the LS-DAG) A system of triangular structural equations corresponding to Example 7.3.1 is

$$\begin{aligned} Y &:= f_Y(D, X, \epsilon_Y), \\ D &:= f_D(Z, X, \epsilon_D), \\ X &:= \epsilon_X, \\ Z &:= \epsilon_Z, \end{aligned}$$

where ϵ_Y, ϵ_X, ϵ_D, and ϵ_Z are mutually independent.

Factual distributions in DAG models have a beautiful Markov factorization structure, which allows for a simple representation of the joint distribution of all variables.

Example 7.3.3 (Factual Law in LS-DAG) Noting the dependences of each variable in the LS-DAG, we can write the joint distribution (density) p of Y, D, X, Z as

$$\mathsf{p}(y, d, x, z) = \mathsf{p}(y|d, x)\,\mathsf{p}(d|x, z)\,\mathsf{p}(x)\,\mathsf{p}(z).$$

Indeed,

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

by Bayes' law. Then $\mathsf{p}(y|d, x, z) = \mathsf{p}(y|d, x)$ as the distribution

of Y is independent of Z, given its parents D and X. Further, $\mathsf{p}(d,x,z) = \mathsf{p}(d|x,z)\,\mathsf{p}(x,z)$, by Bayes' law, and $\mathsf{p}(x,z) = \mathsf{p}(z)\,\mathsf{p}(x)$ by independence.

General DAGs

The purpose of the rest of this section is to give concise general definitions.

A graph G is an ordered pair (V,E), where $V = \{1,...,J\}$ is a collection of vertices/nodes and E is a matrix of edges $e_{ij} \in \{0,1\}$ – that is, $E = \{e_{ij} : (i,j) \in V^2\}$.

Given a collection of random variables $X = (X_j)_{j\in V}$, we associate each index j with the name "X_j" whenever convenient. If the edge (i,j) is present, namely $e_{ij} = 1$, we read it as

$$\text{"}X_i \to X_j\text{" or "}X_i \text{ is an immediate cause of } X_j\text{."}$$

Consider a strict partial order $<$ on V induced by E, where $X_j < X_k$ (we read this as "X_j is determined before X_k") means that either $X_j \to X_k$ or $X_j \to X_{v_1} \to ...X_{v_m} \to X_k$ is true for some v_ℓ's in V. A partial ordering of V exists if for each j the statement $X_j < X_j$ is not true. [6] Note that we may interchangeably use random variable names, X_ℓ, or their indices ,ℓ, when referring to nodes in the graph.

6: The latter statement means that there are no *cycles*.

Definition 7.3.1 (DAG) *The graph $\mathsf{G} = (V,E)$ is a DAG if the graph has no cycles, that is, if V is partially ordered by the edge structure E.*

Example 7.3.4 (LS-DAG continued) In our example (Example 7.3.1), we had vertices $V = \{1,2,3,4\}$ identified with Y, D, X, Z, and the edge set

$$E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

The partial ordering is $X < D$, $X < Y$, $Z < D$, $D < Y$.

Definition 7.3.2 (Parents, Ancestors, Descendants on a DAG) *The parents of X_j are the set $Pa_j := \{X_k : X_k \to X_j\}$. The children*

of X_j are the set $Ch_j := \{X_k : X_j \to X_k\}$. The ancestors of X_j are the set $An_j := \{X_k : X_k < X_j\} \cup \{X_j\}$. The descendants of X_j are the set $Ds_j := \{X_k : X_k > X_j\}$.

Definition 7.3.3 (Paths and Backdoor Paths on DAGs) *A directed path is a sequence $X_{v_1} \to X_{v_2} \to ...X_{v_m}$. A non-directed path is a path, where some arrows (but not all) arrows are replaced by $\leftarrow$. A collider node is a node X_j such that $\to X_j \leftarrow$. A backdoor path from X_l to X_k is an undirected path that starts at X_l and ends with an incoming arrow $\to X_k$.*

From DAGs to ASEMs

Every causal DAG implicitly defines a nonparametric acyclic structural equation model. Thus the two objects are simply different representations or views of the same assumptions on the data generating process and the stochastic potential or counterfactual outcome processes. DAGs are simply a visual depiction of ASEMs and ASEMs are simply a structural equation based expression of DAGs.

Definition 7.3.4 (ASEM) *The ASEM corresponding to the DAG $\mathsf{G} = (V, E)$ is the collection of random variables $\{X_j\}_{j \in V}$ such that*

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

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

Definition 7.3.5 (Linear ASEM) *The linear ASEM is an ASEM where the equations are linear:*

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

here we identify functions $\{f_j\}$ with coefficient vectors $\{f_j\}$.

In linear ASEMs we may replace the requirement of independent errors by the weaker requirement of uncorrelated errors.

Definition 7.3.6 (Structural/Potential Response Processes) *The structural potential response processes for the ASEM corresponding to the DAG $\mathsf{G} = (V, E)$ are given by the structure:*

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

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

Definition 7.3.7 (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, and the variables $\{X_j\}_{j\in V}$ are called endogenous variables. Endogenous variables are determined by the model equations, while exogenous variables are not.

The joint distribution of variables in ASEMs is generally characterized as follows:

Theorem 7.3.1 (Factual Law via Markovian Factorization) *The general ASEM model, given by $(X_j)_{j\in V}$ with an associated DAG $\mathsf{G}(V, E)$, obeys the following equivalent properties:*

- ***Factorization**: The law admits factorization:*

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

- ***Local Markov Property**: All variables are independent of their non-descendants given their parents.*

Counterfactuals Induced by Interventions

We next discuss counterfactuals generated by interventions. We first consider counterfactuals in the Less Simple DAG example (Example 7.3.1). Note that we use the abbreviation "CF" to denote "counterfactual."

Example 7.3.5 (CF-ASEM Induced by Do for LS-DAG Example) Consider the ASEM from Example 7.3.1. A counterfactual system induced by $do(D = d)$ is

$$\begin{aligned} Y(d) &:= f_Y(X, d, \epsilon_Y), \\ d&, \\ Z &= \epsilon_Z, \\ X &= \epsilon_X, \end{aligned}$$

where $\epsilon_X, \epsilon_Z, \epsilon_Y$ are mutually independent. The corresponding graph, provided in Figure 7.7, is denoted by $\mathsf{G}(d)$.

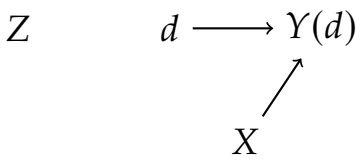

Figure 7.7: CF LS-DAG induced by $do(D = d)$ intervention.

Example 7.3.6 (CF-ASEM Induced by Fix for LSDAG Example) Consider the ASEM from Example 7.3.1. A counterfactual SEM induced by fix$(D = d)$ takes the following form:

$$\begin{aligned} Y(d) &:= f_Y(X, d, \epsilon_Y), \\ d, & \\ D &:= f_D(X, Z, \epsilon_D), \\ Z &:= \epsilon_Z, \\ X &:= \epsilon_X, \end{aligned}$$

where $\epsilon_X, \epsilon_Z, \epsilon_D, \epsilon_Y$ are mutually independent. The corresponding graph, provided in Figure 7.8, is denoted by $\widetilde{\mathsf{G}}(d)$.

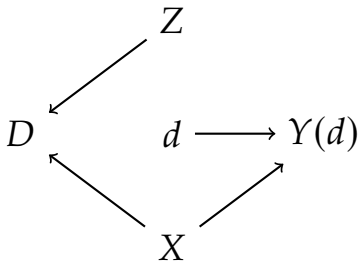

Figure 7.8: CF LS-DAG (SWIG) induced by the $fix_Y(D = d)$ intervention.

We now give a more general definition.

Definition 7.3.8 (Counterfactual ASEM induced by Do Intervention) *The intervention $do(X_j = x_j)$ on an ASEM is said to create the CF-ASEM defined by the modified graph*

$$\mathsf{G}(x_j) = (V, E^*)$$

and collection of counterfactual variables

$$(X_k^*)_{k \in V}$$

where

- *the edges incoming to the node j are set to zero, namely $e_{ij}^* = 0$ for all i,*
- *the remaining edges are preserved, namely $e_{ik}^* = e_{ik}$, for all i and $k \neq j$, and*
- *the counterfactual random variables are defined as*

$$\begin{aligned} X_k^* &:= f_k(Pa_k^*, \epsilon_k), \text{ for } k \neq j. \\ X_j^* &:= x_j \end{aligned}$$

where Pa_k^* are parents of X_k^* $(k \neq j)$ under E^*.

The do intervention modifies the graph G to $\mathsf{G}(x_j)$ by removing edges. Pearl [10] has described this process as "surgery."[7] We next define the do notation to mean

$$\Big((X_\ell)_{\ell \in V} : \mathrm{do}(x_j)\Big) := (X_\ell^*)_{\ell \in V}.$$

7: This sounds a bit painful.

Definition 7.3.9 (Counterfactual ASEM induced by Fix Intervention) *The intervention* $\text{fix}(X_j = x_j)$ *on an ASEM is said to create the CF-ASEM defined by the modified SWIG*

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

and collection of counterfactual variables

$$(X_k^*)_{k \in V} \cup (X_a^*)$$

where we split the node X_j *into* $X_j^* := X_j$ *and the new deterministic node* a

$$X_a^* := x_j,$$

where

- *the node* X_a *inherits only outgoing edges from* X_j *and no incoming edges; namely* $\tilde{e}_{ai} = e_{ji}$ *for all* i *and* $\tilde{e}_{ia} = 0$ *for all* i*;*
- *the node* X_j^* *inherits only incoming edges from* X_j *and no outgoing edges, namely* $\tilde{e}_{ij} = e_{ij}$ *for all* i *and* $\tilde{e}_{ji} = 0$ *for all* i*;*
- *all the remaining edges are preserved, namely* $\tilde{e}_{ik} = e_{ik}$*, for all* i *and* $k \neq j$ *and* $k \neq a$*; and*
- *the counterfactual random variables are assigned according to*

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

  where Pa_k^* *are parents of* X_k^* $(k \neq j)$ *under* $\tilde{E}$*.*

Intervention induces new counterfactual distributions for the endogenous variables; see Appendix 7.A for details.

7.4 Testable Restrictions and d-Separation

Next we examine the constraints on the data generating process that are implied by a given DAG.

For this we turn to a fundamental theorem in DAGs. We will define the concept of d-separation and prove that d-separation implies conditional independence. This property is typically referred to as a global Markov condition that is implied by the DAG. In order to define this property, we need a few more definitions.

The "d" here denotes "directional" as the direction of arrows in a DAG is important for understanding conditional independence relations; see, e.g., Pearl [10] Chapter 11.

Definition 7.4.1 (Blocked Paths) *A path π is is said to be blocked by a subset of nodes S if and only if*

(a) π contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that m is in S;
(b) Or, π contains a collider $i \to m \leftarrow j$, where neither m nor any descendant of m is in S.

A path that is not blocked is called open.

In Figure 7.9 the (backdoor) path $Y \leftarrow X \to D$ is blocked by $S = X$.

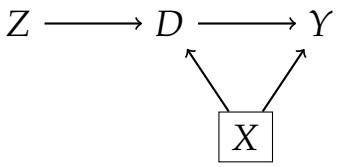

Figure 7.9: The path $Y \leftarrow X \to D$ is blocked by conditioning on X.

The following definition allows empty sets as conditioning sets.

Definition 7.4.2 (Opening a Path by Conditioning) *A path containing a collider is opened by conditioning on it or its descendant.*

In Figure 7.10 the path $Y \to C \leftarrow D$ is blocked, but becomes open by conditioning on the collider $S = C$.

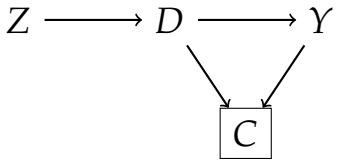

Figure 7.10: The path $Y \to C \leftarrow D$ is blocked, but becomes open by conditioning on C.

The following defines a key graphical property of DAG, which can be used to deduce key statistical independence restrictions.

Definition 7.4.3 (d-Separation) *Given a DAG G, a set of nodes S d-separates nodes X and Y if nodes in S block all paths between X and Y. d-separation is denoted as*

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

The following is a fundamental result concerning the conditional independence relations encoded in the graphs.

Theorem 7.4.1 (Verma and Pearl [11]; Conditional Independence from d-Separation) *d-Separation implies conditional independence:*

- ***Global Markov:*** $(Y \perp\!\!\!\perp_d X \mid S)_{\mathsf{G}} \implies Y \perp\!\!\!\perp X \mid S$.

Figuratively speaking, conditioning on S breaks the information flow between Y and X, meaning that Y can't be predicted by X, conditional on S, and vice versa.

This fundamental result is very intuitive and can be verified directly in simple examples. However, the formal proof is

difficult. The reverse implication is not true in general, but is argued to hold "generically" as we discuss in Section 7.5.

Example 7.4.1 We show a couple of examples illustrating that d-separation implies conditional independence:

1. In Figure 7.11, the variables X and Y are d-separated by $S = (Z, U)$, because S blocks all paths between X and Y. We also have Y is independent of X conditional on S: By the Markov factorization property, $\mathsf{p}(y, x \mid u, z) = \mathsf{p}(y \mid x, z, u)\,\mathsf{p}(x \mid z, u) = \mathsf{p}(y \mid u, z)\,\mathsf{p}(x \mid z, u)$. This equality provides a testable restriction.

2. In Figure 7.12, the variables X and Y are d-separated by $S = Z$, because S blocks all paths between X and Y. We also have Y is independent of X conditional on S: By the Markov factorization property, $\mathsf{p}(y, x \mid z) = \mathsf{p}(y \mid z)\,\mathsf{p}(x \mid z)$. This equality provides a testable restriction.

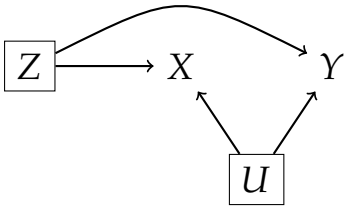

Figure 7.11: Example of d-separation.

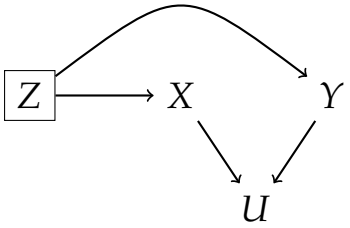

Figure 7.12: Example of d-separation.

These testable restrictions are called exclusion restrictions in econometrics because

$$Y \perp\!\!\!\perp X \mid Z \text{ is equivalent to } \mathsf{p}(y \mid x, z) = p(y \mid z), \tag{7.4.1}$$

where the equivalence follows from Bayes' law. In particular,

$$\mathrm{E}[g(Y) \mid X, Z] = \mathrm{E}[g(Y) \mid Z] \tag{7.4.2}$$

for any bounded function g of Y. (7.4.2) means that X is excluded from the best predictor of $g(Y)$ using X and Z. There are many tests of such restrictions available in the literature.[8] Perhaps one of the reasons for which there are many such tests is that conditional independence testing is formally impossible; see [12]. In practice, the formal impossibility means that any test must be carefully crafted to target specific features within a statistical model as no generic, uniformly valid testing procedure exists.

8: E.g, the reader can search Google Scholar for conditional independence tests, exclusion restrictions tests, or conditional moment tests.

With specific structure provided, conditional independence testing can be relatively straightforward. For example, it reduces to testing hypotheses about linear regression coefficients within a linear ASEM.

Implementation of Tests in Linear ASEMs. Consider the hypothesis that Y is independent of X, given Z. In linear ASEMs, we can test this hypothesis by testing whether the

coefficient $\alpha = 0$ in the projection equation

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

We can perform this test easily with the tools we've developed so far. See R: Dagitty Notebook and Python: Pgmpy Notebook for an example.

Such tests are of course available under structures that are more general than in linear model. For example, [12] exploits debiased ML ideas (introduced in Chapter 4 and further developed in Chapter 10 in this text) to set up testing of exclusion restrictions in some nonlinear models.

Remark 7.4.1 (Equivalence of Local and Global Markov Properties) The local Markov property, the Markov factorization, and the global Markov property are equivalent (Pearl [10]). Therefore, one can use any of these properties to set up tests of the validity of the Markov structure.

7.5 Falsifiability and Causal Discovery*

Here, we provide a brief discussion of whether it is possible to falsify (reject) a causal structure encoded by a DAG with data.

Equivalence Classes and Falsifiability

Definition 7.5.1 (Equivalence Classes) *The class of DAGs that induce the same joint distribution of variables is called an equivalence class, and members of an equivalence class may be described as Markov equivalent. DAGs that produce the same joint distribution variables cannot be distinguished from each other.*

Pearl [10] shows that the equivalence class of a DAG is given by reversing any edges such that any such reversal does not destroy existing or create new *v-structures*: converging arrows whose tails are not connected by an edge.

The equivalence classes of a DAG are called PDAGs (partially directed acyclic graphs). We plot them by erasing arrowheads that can be oriented in the opposite direction without adding or removing v-structures. We illustrate PDAGs in Figures 7.13 and 7.14.

Figure 7.13 starts with the triangular structural equation model from Example 7.2.1. Figure (a) is the original DAG implied by the model. To produce the PDAG, shown in (b), we consider reversing each of the arrows from X to Y, X to P, and P to Y. Because each of the nodes is connected, there are no v-structures in the original DAG, and there is similarly no possible reversal that could add a v-structure. As such, the PDAG is simply the original DAG with all arrows removed. In this case, the DAG structure produces no testable implications.

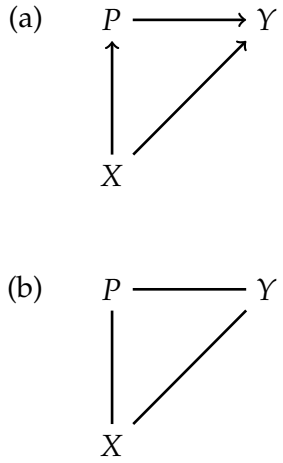

Figure 7.13: The original DAG, (a), and the equivalence class or PDAG, (b), for the TSEM example, Example 7.2.1. The undirected edges in the PDAG mean that they can be directed in any direction as long as this does not create a cycle. In empirical analysis directionality must therefore be deduced and assumed from the context.

Figure 7.14 starts from a more elaborate DAG than the simple TSEM. We refer to this DAG as "Pearl's Example" because it shows up repeatedly as an illustration in Pearl's work; see, e.g., [2]. Figure (a) is the original DAG defining the model. We produce the PDAG in (b) by considering the reversal of all combinations of arrows connecting the eight nodes. Here, there are only two reversals, changing $Z_2 \to X_3$ to $Z_2 \leftarrow X_3$ and changing $Z_1 \to X_1$ to $Z_1 \leftarrow X_1$, that do not destroy any existing v-structures or create new v-structures. For example, reversing the arrow $Z_2 \to X_2$ would destroy the v-structure $Z_2 \to X_2$ and $Z_1 \to X_2$. As such, the PDAG in (b) is almost identical to the DAG in (a) with the exception that the arrows between Z_2 and X_3 and between Z_1 and X_1 have been removed. In this case, the DAG encodes a model which includes exclusion restrictions or testable implications and is potentially falsifiable.

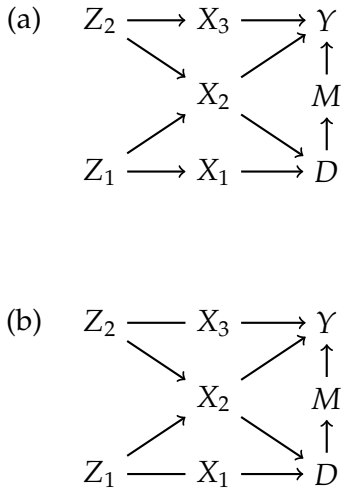

Figure 7.14: The original DAG, (a), and the equivalence class or PDAG, (b), for the Pearl's Example. The undirected edges in the PDAG mean that they can be directed in any direction as long as this does not create a cycle. Only two edges can be reoriented here.

Remark 7.5.1 (Falsifiability) The edge matrix E of a graph is *triangular* if rows of E can be rearranged to have only 1's below the diagonal. In the absence of any further restrictions, an ASEM with graph $\mathsf{G} = (V, E)$ has testable implications if E is not triangular. If E is triangular, then any law p of any arbitrary collection of random variables $(X_j)_{j \in V}$ indexed by V can be factorized as

$$\mathsf{p}(\{x\}_{j \in V}) = \prod_{j \in V} \mathsf{p}(x_j \mid pa_j).$$

With population data we have p and can check if it factorizes according to V. If matrix E is triangular, p always obeys the factorization property. This is to say that there are no exclusion restrictions in the model.

Example 7.5.1 (TSEM continued) In the TSEM example (Example 7.2.1, we have vertices $V = \{1, 2, 3\}$ identified with

Y, P, X and the "triangular" edge set

$$E = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

In the absence of other assumptions, the corresponding TSEM implies no falsifiable restrictions. The equivalence class of the DAG model for this case is generated by rearranging the rows of E in 3! ways, which is equivalent to rearranging the names (Y, P, X) for the nodes.

Example 7.5.2 (Pearl's Example) The DAG given in Figure 7.14 has vertices $V = \{1, \ldots, 8\}$ identified with Y, M, D, X_1, X_2, X_3, Z_1 ,Z_2 and the edge set

$$E = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

This edge set cannot be rearranged to have only ones below the diagonal. The DAG in this case has testable implications, and the equivalence class of the DAG model can only involve changing edges between Z_1 and X_2 and between Z_2 and X_3.

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

$$Y \perp\!\!\!\perp X|S \implies (Y \perp\!\!\!\perp_d X|S)_{\mathsf{G}} \tag{7.5.1}$$

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 (7.5.1) holds, we are said to impose the *faithfulness* assumption on p.

Example 7.5.3 (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 (7.5.1) 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 7.5.2 (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 [13] for an explicit causal discovery algorithm and the review provided in [14]. 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-dimensional graphs, which creates difficulty in inferring the DAG structure from an estimated version $\hat{\mathsf{p}}$.

Example 7.5.4 (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 $\alpha = 0$, $\alpha = .1$,

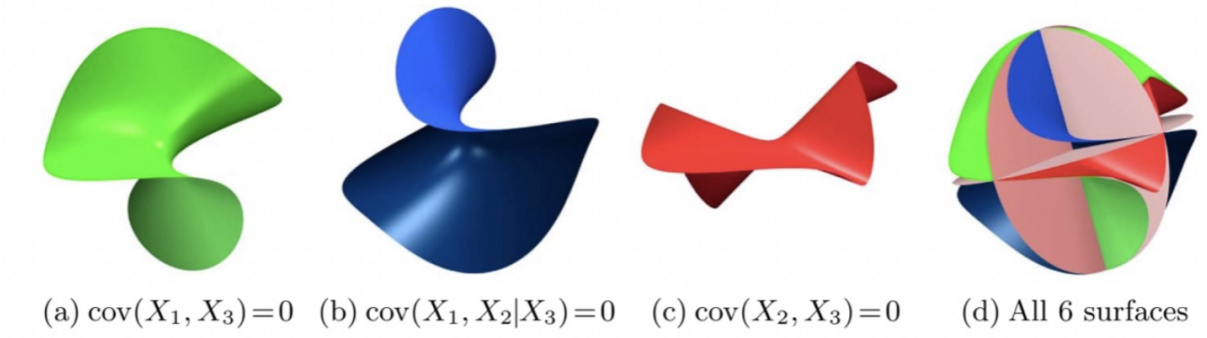

Figure 7.15: Uhler et. al [15]: A set of "unfaithful" distributions p in the simple triangular Gaussian SEM/DAG: $X_1 \to X_2, (X_1, X_2) \to X_3$. The set is parameterized in terms of the covariance of (X_1, X_2, X_3). The right panel shows the set of unfaithful distributions, and the three other panels 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.

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.[15]'s analysis shows.

See Uhler et al's [15] figure; reproduced in Figure 7.15.

The observations above motivate a form of sensitivity analysis – e.g., Conley et al. [16] – 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.

Notebooks

- R: Dagitty Notebook employs the R package "dagitty" to analyze Pearl's Example (introduced in Figure 7.14) 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.

- R: Dosearch Notebook employs the R-package "dosearch" to analyze Pearl's Example (introduced in Figure 7.14). 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.

Additional resources

- Dagitty.Net is an excellent online resource where you can plot and analyze causal DAG models online. It contains many interesting examples of DAGs used in empirical analysis in various fields.

- Causalfusion.Net is another excellent online resource where you can plot and analyze causal DAG models. This resource covers many different deviations from the standard framework.

Study Problems

The study problems ask learners to analyze Pearl's Example (introduced in Figure 7.14). The provided notebooks are a useful starting point for answering these questions.

Recall that Pearl's Example is structured as follows:

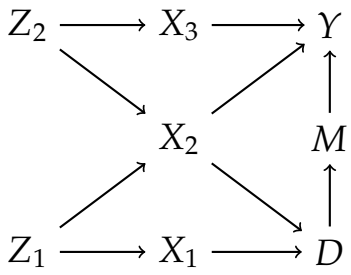

Figure 7.16: Pearl's Example

1. Consider Pearl's Example and answer the following questions. The best way to answer this question is to use computational packages (but please explain the principles the package is using).
   a) What are the testable implications of the assumptions embedded in the model? Hint: The testable implications are derived from the d-separation criterion.
   b) Assume that only variables D, Y, X_2 and M are measured, are there any testable implications?
   c) Now assume only D, Y, and X_2 are measured. Are there any testable implications?
   d) Now assume that all of the variables but X_2 (7 in total) are measured. Are there any testable restrictions?
   e) Assume that an alternative model, competing with Model 1, has the same structure, but with the $X_2 \rightarrow D$ arrow reversed. What statistical test would distinguish between the two models?
2. Work through the proof that d-separation implies conditional independence in Section 7.C. Supply the steps of the proof that were left as a homework or reading exercise.

7.A Counterfactual Distributions*

Interventions induce new counterfactual distributions for endogenous variables. We can readily compute these distributions from the definitions of interventions, as illustrated in the following for the do intervention.

Example 7.A.1 (Counterfactual Law for Do Intervention in LS-DAG (Example 7.3.1)) We can write the counterfactual distribution of $Y(d), Z, X$ in terms of the factual distribution as

$$\mathsf{p}(y, z, x : \mathrm{do}(d)) = \mathsf{p}(y|d, x)\,\mathsf{p}(z)\,\mathsf{p}(x).$$

Indeed,

$$\mathsf{p}(y, z, x : \mathrm{do}(d)) = \mathsf{p}(y|z, x : do(d))\,\mathsf{p}(z, x : do(d)),$$

by definition and Bayes' law. We also have $\mathsf{p}(y|z, x : do(d)) = \mathsf{p}(y|d, x)$ and $\mathsf{p}(z, x : do(d)) = \mathsf{p}(z, x)$ by the definition of the counterfactual ASEM, and $\mathsf{p}(z, x) = \mathsf{p}(z)\,\mathsf{p}(x)$ by indepen-

dence of Z and X.

Theorem 7.A.1 (Counterfactual Law Induced by the Do Intervention) *The induced law p_{X^*} of the counterfactual variables $X^* = (X^*_\ell)_{\ell \in V \setminus j}$ induced by $do(X_j = x_j)$ can be stated in terms of the factual law as follows:*

$$\mathsf{p}(\{x_\ell\}_{\ell \in V \setminus j} : \mathrm{do}(x_j)) := \mathsf{p}_{X^*}(\{x\}_{\ell \in V \setminus j}) = \prod_{\ell \in V \setminus j} \mathsf{p}(x_\ell \mid pa^*_\ell),$$

where $\{x\}_{\ell \in V \setminus j}$ denotes the point where the density function is evaluated, pa^*_j denotes the parental values under the new edge structure, and p denotes the factual law.

The result follows immediately from the Markov factorization property and the definition of counterfactuals under the do intervention. This characterization is interesting in its own right, because it can be used for identification and inference on the counterfactual laws directly, provided that we are willing to model the distribution of the variables. The use of Bayesian methods can be fruitful for this purpose.

These type of formulas are often called "g-formulas" and first appeared in the work [17] of James Robins in 1986 (using another "tree-based" form of causal graphs).

7.B Review of Conditional Independence

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

Lemma 7.B.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.C Theoretical Details of d-Separation⋆

Here we explain why d-separation implies conditional independence.[9]

Lemma 7.C.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* $X' < X$ *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 7.17):

- For any $W \in \mathsf{X} \cup \mathsf{Z}_1$, $Pa_W \subseteq \mathsf{X} \cup \mathsf{Z}$;[10]
- For any $W \in \mathsf{Y} \cup \mathsf{Z}_2$, $Pa_W \subseteq \mathsf{Y} \cup \mathsf{Z}$.[11]

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.[12] It follows that $\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}$.[13] □

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 7.4.1.

9: 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.

10: 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.

11: 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.

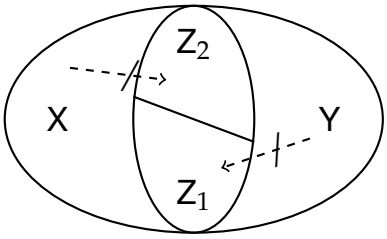

Figure 7.17: Pictorial representation of key argument in Lemma 7.C.1.

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

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

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.

Given the key claim, Lemma 7.C.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.[14] □

14: 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.B.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',[15] 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:

15: 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. □

Bibliography

[1] Judea Pearl and Dana Mackenzie. *The Book of Why*. Penguin Books, 2019 (cited on page 169).

[2] Judea Pearl. 'Causal diagrams for empirical research'. In: *Biometrika* 82.4 (1995), pp. 669–688 (cited on pages 170, 186).

[3] Trygve Haavelmo. 'The probability approach in econometrics'. In: *Econometrica* 12 (1944), pp. iii–vi+1–115 (cited on pages 170, 172).

[4] James Heckman and Rodrigo Pinto. 'Causal analysis after Haavelmo'. In: *Econometric Theory* 31.1 (2015 (NBER 2013)), pp. 115–151 (cited on pages 170, 175).

[5] 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 170, 175).

[6] Alfred Marshall. *Principles of Economics: Unabridged Eighth Edition*. Cosimo, Inc., 2009 (cited on page 173).

[7] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. New York: The Macmillan company, 1928 (cited on page 175).

[8] Frederick Eberhardt and Richard Scheines. 'Interventions and causal inference'. In: *Philosophy of Science* 74.5 (2007), pp. 981–995 (cited on page 175).

[9] Juan Correa and Elias Bareinboim. 'General Transportability of Soft Interventions: Completeness Results'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 10902–10912 (cited on page 175).

[10] Judea Pearl. *Causality*. Cambridge University Press, 2009 (cited on pages 176, 181, 182, 185).

[11] Thomas Verma and Judea Pearl. *Influence diagrams and d-separation*. Tech. rep. Cognitive Systems Laboratory, Computer Science Department, UCLA, 1988 (cited on page 183).

[12] 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 pages 184, 185).

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

[14] 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).

[15] 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).

[16] 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).

[17] James Robins. 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect'. In: *Mathematical Modelling* 7.9-12 (1986), pp. 1393–1512 (cited on page 192).