17

Directed Graphs and Conditional Independence

17.1 Introduction

A directed graph consists of a set of nodes with arrows between some nodes. An example is shown in Figure 17.1.

Graphs are useful for representing independence relations between variables. They can also be used as an alternative to counterfactuals to represent causal relationships. Some people use the phrase **Bayesian network** to refer to a directed graph endowed with a probability distribution. This is a poor choice of terminology. Statistical inference for directed graphs can be performed using

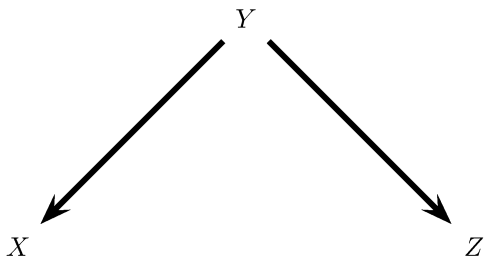


FIGURE 17.1. A directed graph with vertices $V = \{X, Y, Z\}$ and edges $E = \{(Y, X), (Y, Z)\}$.

frequentist or Bayesian methods, so it is misleading to call them Bayesian networks.

Before getting into details about directed acyclic graphs (DAGs), we need to discuss conditional independence.

17.2 Conditional Independence

17.1 Definition. *Let X, Y and Z be random variables. X and Y are* **conditionally independent given** *Z, written $X \amalg Y \mid Z$, if*

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z). \tag{17.1}$$

for all x, y and z.

Intuitively, this means that, once you know Z, Y provides no extra information about X. An equivalent definition is that

$$f(x|y,z) = f(x|z). \tag{17.2}$$

The conditional independence relation satisfies some basic properties.

17.2 Theorem. *The following implications hold:* [1]

$$\begin{aligned}
X \amalg Y \mid Z &\implies Y \amalg X \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies U \amalg Y \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies X \amalg Y \mid (Z,U) \\
X \amalg Y \mid Z \text{ and } X \amalg W \mid (Y,Z) &\implies X \amalg (W,Y) \mid Z \\
X \amalg Y \mid Z \text{ and } X \amalg Z \mid Y &\implies X \amalg (Y,Z).
\end{aligned}$$

17.3 DAGs

A **directed graph** $\mathcal{G}$ consists of a set of vertices V and an edge set E of ordered pairs of vertices. For our purposes, each vertex will correspond to a random variable. If $(X,Y) \in E$ then there is an arrow pointing from X to Y. See Figure 17.1.

[1] The last property requires the assumption that all events have positive probability; the first four do not.

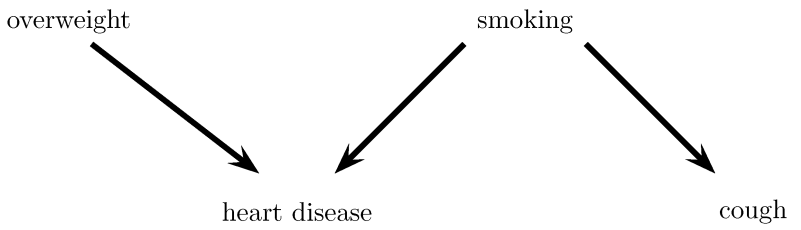
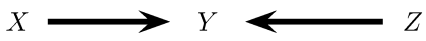


FIGURE 17.2. DAG for Example 17.4.

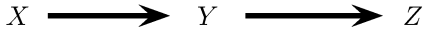
If an arrow connects two variables X and Y (in either direction) we say that X and Y are **adjacent.** If there is an arrow from X to Y then X is a **parent** of Y and Y is a **child** of X. The set of all parents of X is denoted by π_X or $\pi(X)$. A **directed path** between two variables is a set of arrows all pointing in the same direction linking one variable to the other such as:

A sequence of adjacent vertices staring with X and ending with Y but ignoring the direction of the arrows is called an **undirected path.** The sequence $\{X, Y, Z\}$ in Figure 17.1 is an undirected path. X is an **ancestor** of Y if there is a directed path from X to Y (or $X = Y$). We also say that Y is a **descendant** of X.

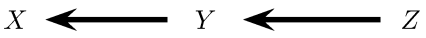
A configuration of the form:



is called a **collider** at Y. A configuration not of that form is called a **non-collider**, for example,



or



The collider property is path dependent. In Figure 17.7, Y is a collider on the path $\{X, Y, Z\}$ but it is a non-collider on the path $\{X, Y, W\}$. When the variables pointing into the collider are not adjacent, we say that the collider is **unshielded**. A directed path that starts and ends at the same variable is called a **cycle.** A directed graph is **acyclic** if it has no cycles. In this case we say that the graph is a **directed acyclic graph** or **DAG**. From now on, we only deal with acyclic graphs.

17.4 Probability and DAGs

Let $\mathcal{G}$ be a DAG with vertices $V = (X_1, \ldots, X_k)$.

17.3 Definition. *If $\mathbb{P}$ is a distribution for V with probability function f, we say that $\mathbb{P}$ is* **Markov to $\mathcal{G}$,** *or that $\mathcal{G}$* **represents** *$\mathbb{P}$, if*

$$f(v) = \prod_{i=1}^{k} f(x_i \mid \pi_i) \tag{17.3}$$

where π_i are the parents of X_i. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

17.4 Example. Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} & f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\ &\quad = f(\text{overweight}) \times f(\text{smoking}) \\ &\quad \times f(\text{heart disease} \,|\, \text{overweight}, \text{smoking}) \\ &\quad \times f(\text{cough} \,|\, \text{smoking}). \quad \blacksquare \end{aligned}$$

17.5 Example. For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function f has the form

$$f(x, y, z, w) = f(x)f(y)f(z \mid x, y)f(w \mid z). \quad \blacksquare$$

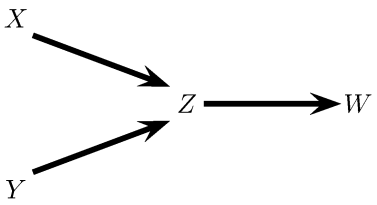


FIGURE 17.3. Another DAG.

The following theorem says that $\mathbb{P} \in M(\mathcal{G})$ if and only if the **Markov Condition** holds. Roughly speaking, the Markov Condition means that every variable W is independent of the "past" given its parents.

17.6 Theorem. *A distribution $\mathbb{P} \in M(\mathcal{G})$ if and only if the following* **Markov Condition** *holds: for every variable W,*

$$W \amalg \widetilde{W} \mid \pi_W \tag{17.4}$$

where $\widetilde{W}$ denotes all the other variables except the parents and descendants of W.

17.7 Example. In Figure 17.3, the Markov Condition implies that

$$X \amalg Y \quad \text{and} \quad W \amalg \{X, Y\} \mid Z. \quad \blacksquare$$

17.8 Example. Consider the DAG in Figure 17.4. In this case probability function must factor like

$$f(a, b, c, d, e) = f(a)f(b|a)f(c|a)f(d|b, c)f(e|d).$$

The Markov Condition implies the following independence relations:

$$D \amalg A \mid \{B, C\}, \quad E \amalg \{A, B, C\} \mid D \quad \text{and} \quad B \amalg C \mid A \quad \blacksquare$$

17.5 More Independence Relations

The Markov Condition allows us to list some independence relations implied by a DAG. These relations might imply other independence relations. Con-

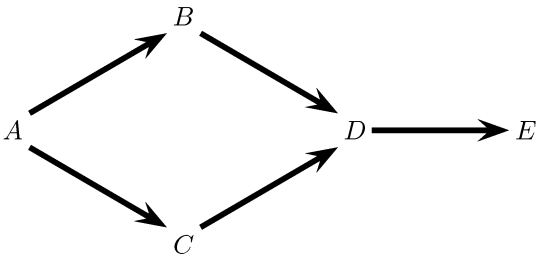


FIGURE 17.4. Yet another DAG.

sider the DAG in Figure 17.5. The Markov Condition implies:

$$X_1 \amalg X_2, \qquad X_2 \amalg \{X_1, X_4\}, \qquad X_3 \amalg X_4 \mid \{X_1, X_2\},$$

$$X_4 \amalg \{X_2, X_3\} \mid X_1, \qquad X_5 \amalg \{X_1, X_2\} \mid \{X_3, X_4\}$$

It turns out (but it is not obvious) that these conditions imply that

$$\{X_4, X_5\} \amalg X_2 \mid \{X_1, X_3\}.$$

How do we find these extra independence relations? The answer is "d-separation" which means "directed separation." d-separation can be summarized by three rules. Consider the four DAG's in Figure 17.6 and the DAG in Figure 17.7. The first 3 DAG's in Figure 17.6 have no colliders. The DAG in the lower right of Figure 17.6 has a collider. The DAG in Figure 17.7 has a collider with a descendant.

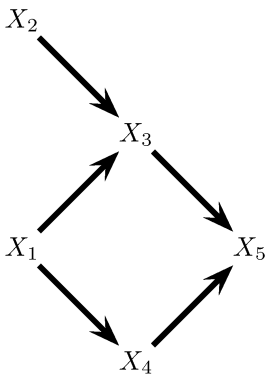


FIGURE 17.5. And yet another DAG.

FIGURE 17.6. The first three DAG's have no colliders. The fourth DAG in the lower right corner has a collider at Y.

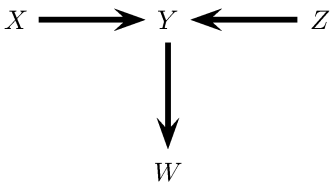


FIGURE 17.7. A collider with a descendant.

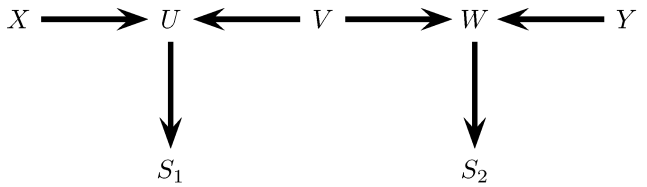


FIGURE 17.8. d-separation explained.

The Rules of d-Separation

Consider the DAGs in Figures 17.6 and 17.7.

1. When Y is not a collider, X and Z are **d-connected**, but they are **d-separated** given Y.
2. If X and Z collide at Y, then X and Z are **d-separated**, but they are **d-connected** given Y.
3. Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 17.7, X and Z are **d-separated** but they are **d-connected** given W.

Here is a more formal definition of d-separation. Let X and Y be distinct vertices and let W be a set of vertices not containing X or Y. Then X and Y are **d-separated given** W if there exists no undirected path U between X and Y such that (i) every collider on U has a descendant in W, and (ii) no other vertex on U is in W. If A, B, and W are distinct sets of vertices and A and B are not empty, then A and B are d-separated given W if for every $X \in A$ and $Y \in B$, X and Y are d-separated given W. Sets of vertices that are not d-separated are said to be d-connected.

17.9 Example. Consider the DAG in Figure 17.8. From the d-separation rules we conclude that:

X and Y are d-separated (given the empty set);
X and Y are d-connected given $\{S_1, S_2\}$;
X and Y are d-separated given $\{S_1, S_2, V\}$.

17.10 Theorem. [2] *Let A, B, and C be disjoint sets of vertices. Then $A \amalg B \mid C$ if and only if A and B are d-separated by C.*

[2] We implicitly assume that $\mathbb{P}$ is **faithful** to $\mathcal{G}$ which means that $\mathbb{P}$ has no extra independence relations other than those logically implied by the Markov Condition.

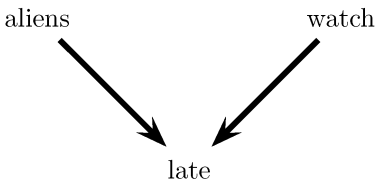


FIGURE 17.9. Jordan's alien example (Example 17.11). Was your friend kidnapped by aliens or did you forget to set your watch?

17.11 Example. The fact that conditioning on a collider creates dependence might not seem intuitive. Here is a whimsical example from Jordan (2004) that makes this idea more palatable. Your friend appears to be late for a meeting with you. There are two explanations: she was abducted by aliens or you forgot to set your watch ahead one hour for daylight savings time. (See Figure 17.9.) Aliens and Watch are blocked by a collider which implies they are marginally independent. This seems reasonable since — before we know anything about your friend being late — we would expect these variables to be independent. We would also expect that $\mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}) > \mathbb{P}(\mathsf{Aliens} = \mathsf{yes})$; learning that your friend is late certainly increases the probability that she was abducted. But when we learn that you forgot to set your watch properly, we would lower the chance that your friend was abducted. Hence, $\mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}) \neq \mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}, \mathsf{Watch} = \mathsf{no})$. Thus, Aliens and Watch are dependent given Late. ■

17.12 Example. Consider the DAG in Figure 17.2. In this example, overweight and smoking are marginally independent but they are dependent given heart disease. ■

Graphs that look different may actually imply the same independence relations. If $\mathcal{G}$ is a DAG, we let $\mathcal{I}(\mathcal{G})$ denote all the independence statements implied by $\mathcal{G}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ for the same variables V are **Markov equivalent** if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. Given a DAG $\mathcal{G}$, let skeleton($\mathcal{G}$) denote the undirected graph obtained by replacing the arrows with undirected edges.

17.13 Theorem. *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if (i)* skeleton($\mathcal{G}_1$) = skeleton($\mathcal{G}_2$) *and (ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders.*

17.14 Example. The first three DAGs in Figure 17.6 are Markov equivalent. The DAG in the lower right of the Figure is not Markov equivalent to the others. ■

17.6 Estimation for DAGs

Two estimation questions arise in the context of DAGs. First, given a DAG $\mathcal{G}$ and data $V_1, \ldots, V_n$ from a distribution f consistent with $\mathcal{G}$, how do we estimate f? Second, given data $V_1, \ldots, V_n$ how do we estimate $\mathcal{G}$? The first question is pure estimation while the second involves model selection. These are very involved topics and are beyond the scope of this book. We will just briefly mention the main ideas.

Typically, one uses some parametric model $f(x|\pi_x; \theta_x)$ for each conditional density. The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(V_i; \theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}|\pi_j; \theta_j),$$

where X_{ij} is the value of X_j for the i^{th} data point and θ_j are the parameters for the j^{th} conditional density. We can then estimate the parameters by maximum likelihood.

To estimate the structure of the DAG itself, we could fit every possible DAG using maximum likelihood and use AIC (or some other method) to choose a DAG. However, there are many possible DAGs so you would need much data for such a method to be reliable. Also, searching through all possible DAGs is a serious computational challenge. Producing a valid, accurate confidence set for the DAG structure would require astronomical sample sizes. If prior information is available about part of the DAG structure, the computational and statistical problems are at least partly ameliorated.

17.7 Bibliographic Remarks

There are a number of texts on DAGs including Edwards (1995) and Jordan (2004). The first use of DAGs for representing causal relationships was by Wright (1934). Modern treatments are contained in Spirtes et al. (2000) and Pearl (2000). Robins et al. (2003) discuss the problems with estimating causal structure from data.

17.8 Appendix

Causation Revisited. We discussed causation in Chapter 16 using the idea of counterfactual random variables. A different approach to causation uses

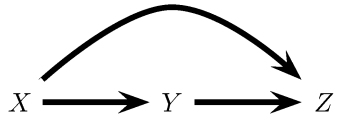


FIGURE 17.10. Conditioning versus intervening.

DAGs. The two approaches are mathematically equivalent though they appear to be quite different. In the DAG approach, the extra element is the idea of **intervention**. Consider the DAG in Figure 17.10.

The probability function for a distribution consistent with this DAG has the form $f(x,y,z) = f(x)f(y|x)f(z|x,y)$. The following is pseudocode for generating from this distribution.

$$\begin{aligned} \text{For } i &= 1,\ldots,n: \\ x_i &\leftarrow p_X(x_i) \\ y_i &\leftarrow p_{Y|X}(y_i|x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y_i) \end{aligned}$$

Suppose we repeat this code many times, yielding data $(x_1,y_1,z_1),\ldots,(x_n,y_n,z_n)$. Among all the times that we observe $Y=y$, how often is $Z=z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$\begin{aligned} \mathbb{P}(Z=z|Y=y) &= \frac{\mathbb{P}(Y=y,Z=z)}{\mathbb{P}(Y=y)} = \frac{f(y,z)}{f(y)} \\ &= \frac{\sum_x f(x,y,z)}{f(y)} = \frac{\sum_x f(x)\,f(y|x)\,f(z|x,y)}{f(y)} \\ &= \sum_x f(z|x,y)\frac{f(y|x)\,f(x)}{f(y)} = \sum_x f(z|x,y)\frac{f(x,y)}{f(y)} \\ &= \sum_x f(z|x,y)\,f(x|y). \end{aligned}$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix Y at the value y. The code now looks like this:

$$\begin{aligned} \text{set } Y &= y \\ \text{for } i &= 1,\ldots,n \\ x_i &\leftarrow p_X(x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y) \end{aligned}$$

Having **set** $Y = y$, how often was $Z = z$? To answer, note that the intervention has changed the joint probability to be

$$f^*(x, z) = f(x)f(z|x, y).$$

The answer to our question is given by the marginal distribution

$$f^*(z) = \sum_x f^*(x, z) = \sum_x f(x)f(z|x, y).$$

We shall denote this as $\mathbb{P}(Z = z|Y := y)$ or $f(z|Y := y)$. We call $\mathbb{P}(Z = z|Y = y)$ **conditioning by observation** or **passive conditioning.** We call $\mathbb{P}(Z = z|Y := y)$ **conditioning by intervention** or **active conditioning.**

Passive conditioning is used to answer a predictive question like:

"Given that Joe smokes, what is the probability he will get lung cancer?"

Active conditioning is used to answer a causal question like:

"If Joe quits smoking, what is the probability he will get lung cancer?"

Consider a pair $(\mathcal{G}, \mathbb{P})$ where $\mathcal{G}$ is a DAG and $\mathbb{P}$ is a distribution for the variables V of the DAG. Let p denote the probability function for $\mathbb{P}$. Consider intervening and fixing a variable X to be equal to x. We represent the intervention by doing two things:

(1) Create a new DAG $\mathcal{G}^*$ by removing all arrows pointing into X;

(2) Create a new distribution $f^*(v) = \mathbb{P}(V = v|X := x)$ by removing the term $f(x|\pi_X)$ from $f(v)$.

The new pair $(\mathcal{G}^*, f^*)$ represents the intervention "set $X = x$."

17.15 Example. You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence $p_{R,W}(r, w) \neq p_R(r)p_W(w)$ where R denotes Rain and W denotes Wet Lawn. Consider the following two DAGs:

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn}.$$

The first DAG implies that $f(w, r) = f(r)f(w|r)$ while the second implies that $f(w, r) = f(w)f(r|w)$ No matter what the joint distribution $f(w, r)$ is, both graphs are correct. Both imply that R and W are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.

Look at the first graph and form the intervention $W = 1$ where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into W

to get the modified graph:

with distribution $f^*(r) = f(r)$. Thus $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ tells us that "wet lawn" does not cause rain.

Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention $W = 1$ on the second graph. There are no arrows into W that need to be broken so the intervention graph is the same as the original graph. Thus $f^*(r) = f(r|w)$ which would imply that changing "wet" changes "rain." Clearly, this is nonsense.

Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.

17.16 Remark. We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable.

We can use DAGs to represent confounding variables. If X is a treatment and Y is an outcome, a confounding variable Z is a variable with arrows into both X and Y; see Figure 17.11. It is easy to check, using the formalism of interventions, that the following facts are true:

In a randomized study, the arrow between Z and X is broken. In this case, even with Z unobserved (represented by enclosing Z in a circle), the causal relationship between X and Y is estimable because it can be shown that $\mathbb{E}(Y|X := x) = \mathbb{E}(Y|X = x)$ which does not involve the unobserved Z. In an observational study, with all confounders observed, we get $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ as in formula (16.7). If Z is unobserved then we cannot estimate the causal effect because $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ involves the unobserved Z. We can't just use X and Y since in this case. $\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|X := x)$ which is just another way of saying that causation is not association.

In fact, we can make a precise connection between DAGs and counterfactuals as follows. Suppose that X and Y are binary. Define the confounding

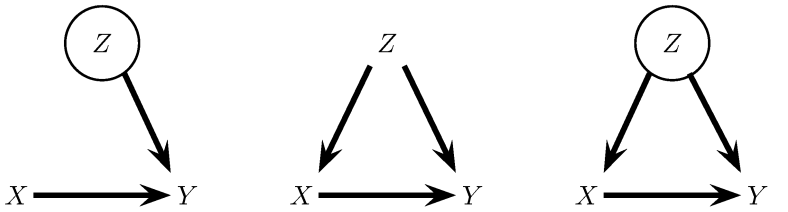


FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

variable Z by

$$Z = \begin{cases} 1 & \text{if } (C_0, C_1) = (0,0) \\ 2 & \text{if } (C_0, C_1) = (0,1) \\ 3 & \text{if } (C_0, C_1) = (1,0) \\ 4 & \text{if } (C_0, C_1) = (1,1). \end{cases}$$

From this, you can make the correspondence between the DAG approach and the counterfactual approach explicit. I leave this for the interested reader.

17.9 Exercises

1. Show that (17.1) and (17.2) are equivalent.

2. Prove Theorem 17.2.

3. Let X, Y and Z have the following joint distribution:

	$Y=0$	$Y=1$
$X=0$	.405	.045
$X=1$	.045	.005

$Z = 0$

	$Y=0$	$Y=1$
$X=0$	.125	.125
$X=1$	.125	.125

$Z = 1$

   (a) Find the conditional distribution of X and Y given $Z = 0$ and the conditional distribution of X and Y given $Z = 1$.

   (b) Show that $X \amalg Y | Z$.

   (c) Find the marginal distribution of X and Y.

   (d) Show that X and Y are not marginally independent.

4. Consider the three DAGs in Figure 17.6 without a collider. Prove that $X \amalg Z | Y$.

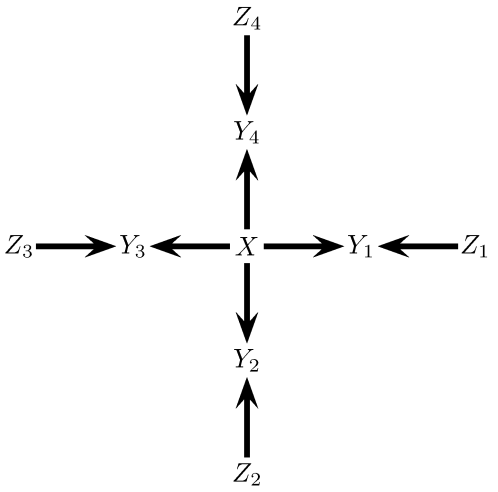


FIGURE 17.12. DAG for exercise 7.

5. Consider the DAG in Figure 17.6 with a collider. Prove that $X \amalg Z$ and that X and Z are dependent given Y.

6. Let $X \in \{0, 1\}$, $Y \in \{0, 1\}$, $Z \in \{0, 1, 2\}$. Suppose the distribution of (X, Y, Z) is Markov to:

$$X \longrightarrow Y \longrightarrow Z$$

   Create a joint distribution $f(x, y, z)$ that is Markov to this DAG. Generate 1000 random vectors from this distribution. Estimate the distribution from the data using maximum likelihood. Compare the estimated distribution to the true distribution. Let $\theta = (\theta_{000}, \theta_{001}, \ldots, \theta_{112})$ where $\theta_{rst} = \mathbb{P}(X = r, Y = s, Z = t)$. Use the bootstrap to get standard errors and 95 percent confidence intervals for these 12 parameters.

7. Consider the DAG in Figure 17.12.

   (a) Write down the factorization of the joint density.

   (b) Prove that $X \amalg Z_j$.

8. Let $V = (X, Y, Z)$ have the following joint distribution

$$X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

$$\begin{aligned} Y \mid X = x &\sim \text{Bernoulli}\left(\frac{e^{4x-2}}{1+e^{4x-2}}\right) \\ Z \mid X = x, Y = y &\sim \text{Bernoulli}\left(\frac{e^{2(x+y)-2}}{1+e^{2(x+y)-2}}\right). \end{aligned}$$

(a) Find an expression for $\mathbb{P}(Z = z \mid Y = y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y = 1)$.

(b) Write a program to simulate the model. Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y = 1)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (a).

(c) (Refers to material in the appendix.) Write down an expression for $\mathbb{P}(Z = 1 \mid Y := y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y := 1)$.

(d) (Refers to material in the appendix.) Modify your program to simulate the intervention "set $Y = 1$." Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y := 1)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (c).

9. This is a continuous, Gaussian version of the last question. Let $V = (X, Y, Z)$ have the following joint distribution

$$\begin{aligned} X &\sim \text{Normal}\,(0, 1) \\ Y \mid X = x &\sim \text{Normal}\,(\alpha x, 1) \\ Z \mid X = x, Y = y &\sim \text{Normal}\,(\beta y + \gamma x, 1). \end{aligned}$$

Here, α, β and γ are fixed parameters. economists refer to models like this as **structural equation models.**

(a) Find an explicit expression for $f(z \mid y)$ and $\mathbb{E}(Z \mid Y = y) = \int z f(z \mid y) dz$.

(b) (Refers to material in the appendix.) Find an explicit expression for $f(z \mid Y := y)$ and then find $\mathbb{E}(Z \mid Y := y) \equiv \int z f(z \mid Y := y) dy$. Compare to (b).

(c) Find the joint distribution of (Y, Z). Find the correlation ρ between Y and Z.

(d) (Refers to material in the appendix.) Suppose that X is not observed and we try to make causal conclusions from the marginal distribution of (Y, Z). (Think of X as unobserved confounding variables.) In particular,

suppose we declare that Y causes Z if $\rho \neq 0$ and we declare that Y does not cause Z if $\rho = 0$. Show that this will lead to erroneous conclusions.

(e) (Refers to material in the appendix.) Suppose we conduct a randomized experiment in which Y is randomly assigned. To be concrete, suppose that

$$\begin{aligned} X &\sim \text{Normal}(0, 1) \\ Y &\sim \text{Normal}(\alpha, 1) \\ Z \mid X = x, Y = y &\sim \text{Normal}(\beta y + \gamma x, 1). \end{aligned}$$

Show that the method in (d) now yields correct conclusions (i.e., $\rho = 0$ if and only if $f(z \mid Y := y)$ does not depend on y).