2 A Tutorial on Learning Causal Influence

Richard E. Neapolitan, Xia Jiang
Northeastern Illinois University
RE-Neapolitan@neiu.edu, xjiang@cbmi.pitt.edu

Abstract

In the 1990's related research in artificial intelligence, cognitive science, and philosophy resulted in a method for learning causal relationships from passive data when we have data on at least four variables. We illustrate the method using a few simple examples. Then we present recent research showing that we can even learn something about causal relationships when we have data on only two variables.

2.1 Introduction

Christensen [1] says "Causality is not something that can be established by data analysis. Establishing causality requires logical arguments that go beyond the realm of numerical manipulation." The argument is that if, for example, we determine that smoking and lung cancer are correlated, we cannot necessarily deduce that smoking causes lung cancer. The following are other causal explanations for smoking and lung cancer being correlated: lung cancer could cause smoking; lung cancer and smoking could cause each other via a causal feedback loop; lung cancer and smoking could have a hidden common cause such as a genetic defect; and lung cancer and smoking could cause some other condition, and we have sampled from a population in which all individuals have this condition (This is called selection bias.). All five causal explanations are shown in Figure 2.1.

This argument that causal relationships cannot be learned from passive data is compelling when we have data on only two variables (e.g. smoking and lung cancer). However, in the 1990's related research in artificial intelligence, cognitive science, and philosophy [6, 11, 13, 17] resulted in a method for learning causal relationships when we have data on at least four variables (or three variables if we have

R. E. Neapolitan and X. Jiang: *A Tutorial on Learning Causal Influence*, StudFuzz **194**, 29–71 (2006)
www.springerlink.com

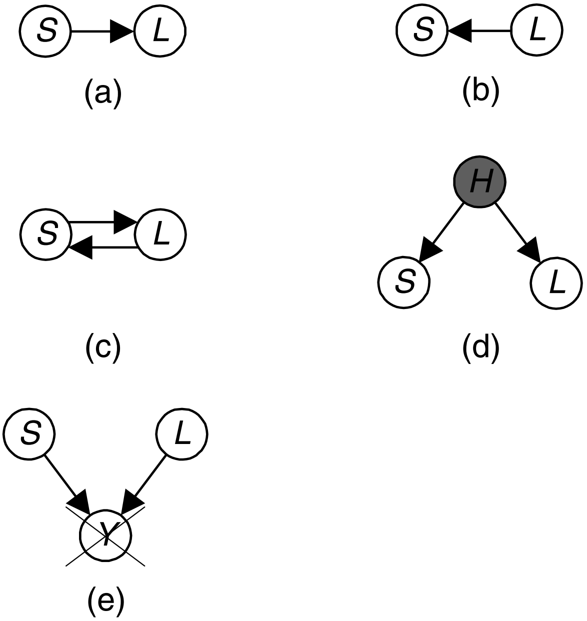


Figure 2.1: All five causal relationships could account for smoking (S) and lung cancer (L) being correlated.

a time ordering of the events). This method is discussed in detail in [11] and [17]. In Section 2 we illustrate the method using a few simple examples. Then in Section 3 we present recent research showing that we can even learn something about causal relationships when we have data on only two variables. In the current section we present some necessary preliminary concepts including a definition of causal networks.

2.1.1 Causation

A common way to learn (perhaps define) causation is via manipulation experiments. We say we **manipulate** X when we force X to take some value, and we say X **causes** Y if there is some manipulation of X that leads to a change in the probability distribution of Y. A manipulation consists of a randomized controlled experiment (**RCE**) using some specific population of entities (e.g. individuals with chest pain) in some specific context (E.g., they currently receive no chest pain medication and they live in a particular geographical area.). The causal relationship discovered is then relative to this population and this context.

Let's discuss how the manipulation proceeds. We first identify the population of entities we wish to consider. Our variables are features of these entities. Next we ascertain the causal relationship we wish to investigate. Suppose we are trying to determine if variable X is a cause of variable Y. We then sample a number of entities from the population. For every entity selected, we manipulate the value of X so that each of its possible values is given to the same number of entities (If X is continuous, we choose the values of X according to a uniform distribution.). After the value of X is set for a given entity, we measure the value of Y for that entity. The more the resultant data shows a dependency between X and Y the more the data supports that X causally influences Y. The manipulation of X can be represented by a variable M that is external to the system being studied. There is one value mi of M for each value xi of X, the probabilities of all values of M are the same, and when M equals mi, X equals xi. That is, the relationship between M and X is deterministic. The data supports that X causally influences Y to the extent that the data indicates $P(yi|mj) \neq P(yi|mk)$ for $j \neq k$. Manipulation is actually a special kind of causal relationship that we assume exists primordially and is within our control so that we can define and discover other causal relationships. Figure 2.2 depicts a manipulation experiment in which we are trying to determine whether smoking causes lung cancer. In that figure, $P(m1) = .5$ means the probability of being selected for smoking is .5, and$P(s1|m1) = 1$ means the probability of smoking ($s1$) is 1 if the person is selected for smoking. Similarly, the $P(s2|m2) = 1$ means the probability of not smoking ($s2$) is 1 if the person is selected for not smoking.

We do not really want to manipulate people and make them smoke. We will see in the next section that it is possible to learn something about whether smoking causes lung cancer from passive data alone.

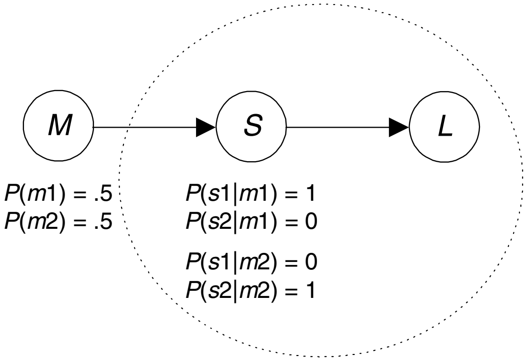


Figure 2.2: A manipulation experiment to determine whether smoking (S) causes lung cancer (L).

First we need to define causal networks.

2.1.2 Causal networks

If we create a causal DAG (directed acyclic graph) $\mathbb{G}$ and assume the observed probability distribution P of the variables/nodes in the DAG satisfies the Markov condition with $\mathbb{G}$, we say we say we are making the **causal Markov assumption**, and $(\mathbb{G}, P)$ is called a **causal network** [10]. A **causal DAG** is a DAG in which there is an edge from X to Y if and only if X is a direct cause of Y. By a 'direct cause' we mean a manipulation of X results in a change in the probability distribution of Y, and there is no subset of variables W of the set of variables in the DAG such that if we knew the values of the variables in W, a manipulation of X would no longer change the probability distribution of Y. A probability distribution P satisfies the **Markov condition** with a DAG $\mathbb{G}$ if the probability of each variable/node in the DAG is independent of its nondescendents conditional on its parents. We will use the notation $I(X, Y|Z)$ to denote that X is independent of Y conditional on Z. That is, $I(X, Y|Z)$ holds if for all values x, y, and

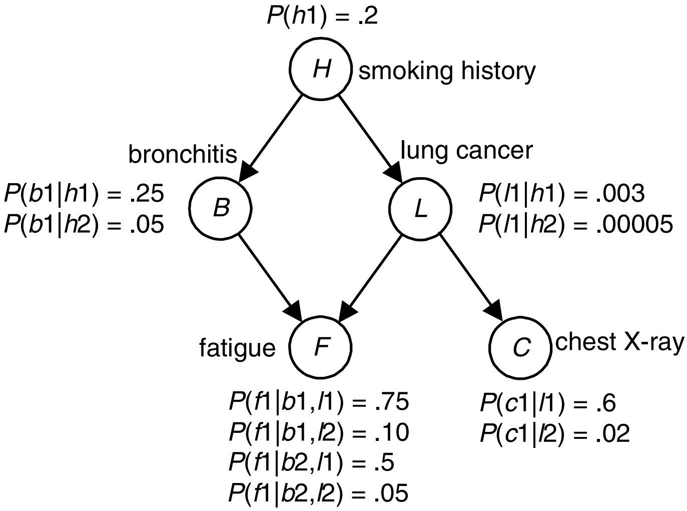


Figure 2.3: A Causal Network

z of X, Y, and Z we have

$$P(x|y,z) = P(x|z).$$

Similarly, if A is a set of variables and X is independent of the variables in A conditional on Z, we write $I(X, \mathsf{A}|Z)$.

Consider the causal network in Figure 2.3. The causal Markov assumption for that network entails the following conditional independencies:

$$\begin{array}{cc} I(B,\{L,C\}|H) & I(F,\{H,C\}|\{L,B\}) \\ I(L,B|H) & I(C,\{H,B,F\}|L). \end{array}$$

Why should we make the causal Markov assumption? A study in [7] supplies experimental evidence for it. We offer the following brief intuitive justification. Given the causal relationship in Figure 2.3, we would not expect bronchitis and lung cancer to be independent because if someone had lung cancer it would make it more probable that they smoked (since smoking is a cause of lung cancer), which would make it more probable that another effect of smoking, namely bronchitis, was present. However, if we knew someone smoked, it would already be more probable that the person had bronchitis. Learning that they had lung cancer could no longer increase the probability of smoking (This probability is now 1.), which means it can't change the probability of bronchitis. That is, the variable H shields B from the influence of L, which is what the causal Markov condition says. Similarly, a positive chest X-ray increases the probability of lung cancer, which in turn increases the probability of smoking, which in turn increases the probability of bronchitis. So a chest X-ray and bronchitis are not independent. However, if we knew the person had lung cancer, the chest X-ray could not change the probability of lung cancer and thereby change the probability of bronchitis. So B is independent of C conditional on L, which is what the causal Markov condition says.

Notice in Figure 2.3 that we show the probability distribution of each variable conditional on the values of its parents. It is a theorem that if P satisfies the Markov condition with $\mathbb{G}$, then P is equal to the product of its conditional distributions in $\mathbb{G}$ (See [10,11] for the proof.). So in a causal network we specify the probability distribution by showing the conditional distributions. If the network is sparse, this is far more efficient than listing every value in the probability distribution. For example, if there are 100 binary variables, there are 2^{100} values in the joint probability distribution. However, if each variable has at most three parents, less than 800 values determine the conditional distributions.

The following conditions must be satisfied for the causal Markov condition to hold:

1. If X causes Y, we must draw an edge from X to Y unless all causal paths from X to Y are mediated by observed variables.

2. There must be no causal feedback loops.

3. There must be no hidden common causes.

4. Selection bias must not be present.

These conditions are discussed in detail in [11]. Perhaps the condition that is most violated is that there can be no hidden common

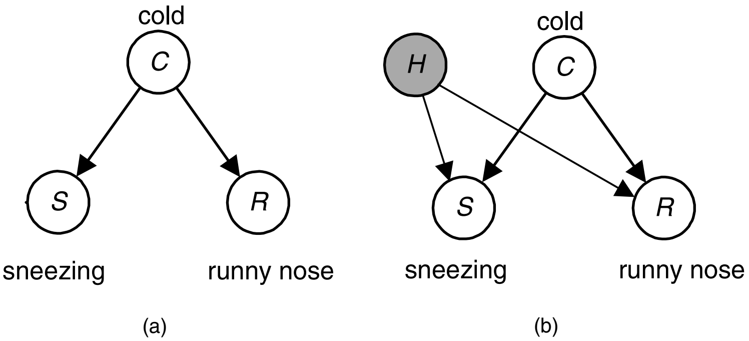


Figure 2.4: The causal Markov assumption would not hold for the DAG in (a) if there is a hidden common cause as depicted in (b).

causes. We discuss that condition further here. Suppose we draw a causal DAG containing the variables cold (C), sneezing (S), and runny nose (R). Then since a cold can cause both sneezing and a runny nose and neither of these conditions can cause each other, we would draw the DAG in Figure 2.4 (a). The causal Markov condition for that DAG would entail $I(S, R|C)$. However, if there were a hidden common cause of S and R as depicted in (b), this conditional independency would not hold because even if the value of C were know, S would change the probability of H, which in turn would change the probability of R. Indeed, there is another cause of sneezing and runny nose, namely hay fever. So when making the causal Markov assumption, we must be certain that we have identified all common causes.

Consider now the causal DAG in Figure 2.5. The pharmaceutical company Merck developed the drug finasteride which lowers dehydrotestosterone (DHT) levels in men, and DHT is believed to be the hormone responsible for baldness in men. However, DHT is also necessary for erectile function, as shown by a study in [7]. Merck wanted to market finasteride as a hair regrowth treatment, but they feared the side effect of erectile dysfunction. So Merck conducted a large scale manipulation study [8], and found that finasteride has a causal effect on DHT level but has no causal effect on erectile dysfunction.

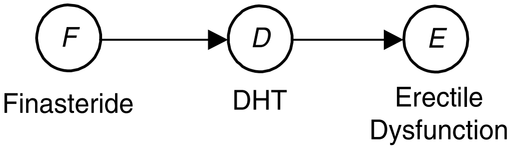


Figure 2.5: Finasteride and erectile function are independent.

How could this be when the causal relationships among finasteride (F), DHT (D), and erectile dysfunction (E) have clearly been found to be those depicted in Figure 2.5? We would expect a causal mediary to transmit an effect from its antecedent to its consequence, but in this case it does not. The explanation is that finasteride cannot lower DHT levels beyond a certain threshold level, and that level is all that is needed for erectile function. So we have $I(F, E)$.

The Markov condition does not entail $I(F, E)$ for the causal DAG in Figure 2.5. It only entails $I(F, E|D)$. When a probability distribution has a conditional independency that is not entailed by the Markov condition, the faithfulness assumption does not hold. If we create a causal DAG $\mathbb{G}$ and assume

1. $(\mathbb{G}, P)$ satisfies the Markov condition,

2. All conditional independencies in the observed distribution P are entailed by the Markov condition in $\mathbb{G}$,

then we say we are making the **causal faithfulness assumption**.

Besides those conditions already stated for the causal Markov assumption to hold, the following additional conditions must be satisfied for the causal faithfulness assumption to hold:

1. We can't have 'unusual' causal relationships as in the finasteride example.

2. We cannot draw an edge from X to Y if for every causal path from X to Y there is a causal mediary in the set of observed variables.

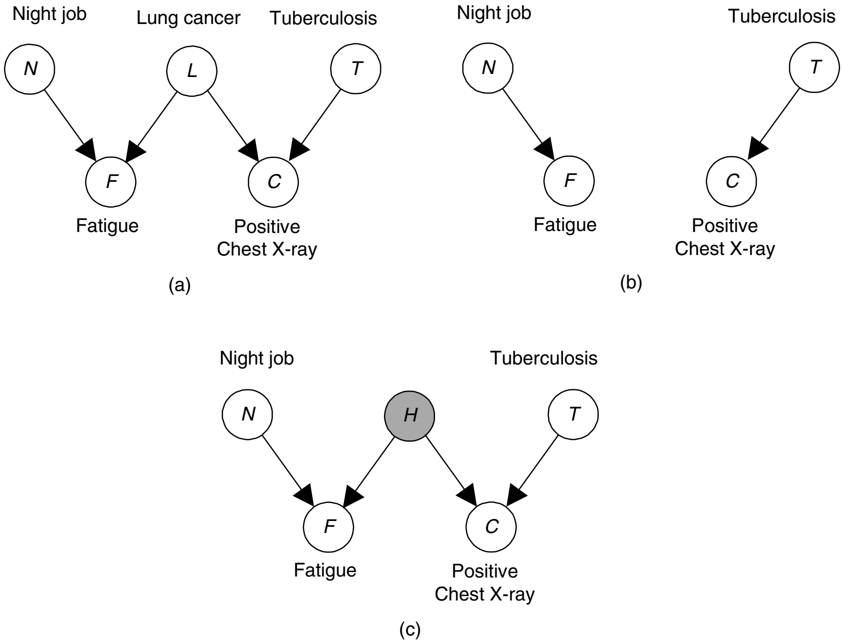


Figure 2.6: If the causal relationships are those shown in (a), P is not faithful to the DAG in (b), but P is embedded faithfully in the DAG in (c).

It seems the main exception to the causal faithfulness assumption (and the causal Markov assumption) is the presence of hidden common causes. Our next assumption eliminates that exception. If we assume the observed probability distribution P of the variables is embedded faithfully in a causal DAG containing the variables, we say we are making the **causal embedded faithfulness assumption**. Suppose we have a probability distribution P of the variables in a set V, V is a subset of W, and $\mathbb{G}$ is a DAG whose set of nodes is W. Then P is **embedded faithfully** in W if all and only the conditional independencies in P are entailed by the Markov condition applied to W and restricted to the nodes in V.

Next we illustrate the causal embedded faithfulness assumption.

Suppose the causal DAG in Figure 2.6 (a) satisfies the causal faithfulness assumption. However, we only observe $\mathsf{V} = \{N, F, C, T\}$. Then the causal DAG containing the observed variables is the one in Figure 2.6 (b). The Markov condition entails $I(F, C)$ for the DAG in Figure 2.6 (b), and this conditional independency is not entailed by the DAG in Figure 2.6 (a). Therefore, the observed distribution $P(\mathsf{V})$ does not satisfy the Markov condition with the causal DAG in Figure 2.6 (b), which means the causal faithfulness assumption is not warranted. However, $P(\mathsf{V})$ is embedded faithfully in the DAG in Figure 2.6 (c). So the causal embedded faithfulness assumption is warranted. Note that this example illustrated a situation in which we identify four variables and two of them have a hidden common cause. That is, we have not identified lung cancer as a feature of humans.

2.2 Learning Causal Influences

Next we show how causal influences can be learned from data if we make either the causal faithfulness or the causal embedded faithfulness assumption. We assume that we have a random sample of entities and know the values of the variables of interest for the entities in the sample. From this sample, we have deduced the conditional independencies among the variables. A method for doing this is described in [11] and [17]. Our confidence in the causal influences we conclude is no greater than our confidence in these conditional independencies.

2.2.1 Making the Causal Faithfulness Assumption

We assume here that the variables satisfy the causal faithfulness assumption and we know the conditional independencies among the variables. Given these assumptions, we present a sequence of examples showing how causal influences can be learned.

Example 1 *Suppose* V *is our set of observed variables,* $\mathsf{V} = \{X, Y\}$, *and our set of conditional independencies is*

$$\{I(X, Y)\}.$$

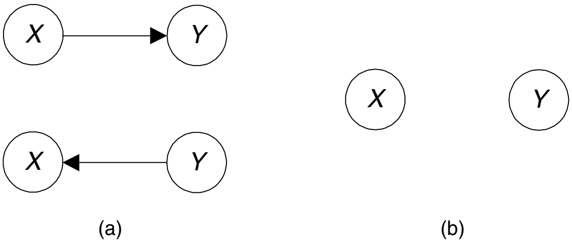


Figure 2.7: If the set of conditional independencies is $\{I(X, Y)\}$, we must have the causal DAG in (b), whereas if it is $\varnothing$, we must have one of the causal DAGs in (a).

Then we cannot have either of the causal DAGs in Figure 2.7 (a). The reason is that the Markov condition, applied to these DAGs, does not entail that X and Y are independent, which means the causal faithfulness assumption is not satisfied. So we must have the causal DAG in Figure 2.7 (b). We conclude X and Y have no causal influence on each other.

Example 2 *Suppose* $\mathsf{V} = \{X, Y\}$ *and our set of conditional independencies is the empty set*

$$\varnothing.$$

Then we cannot have the causal DAG in Figure 2.7 (b). The reason is that the Markov condition, applied to this DAG, entails that X and Y are independent, which means the causal Markov assumption would not be satisfied. So we must have one of the causal DAGs in Figure 2.7 (a). We conclude either X causes Y or Y causes X.

Example 3 *Suppose* $\mathsf{V} = \{X, Y, Z\}$ *and our set of conditional independencies is*

$$\{I(X, Y)\}.$$

Then there can be no edge between X and Y in the causal DAG owing to the reason given in Example 1. Furthermore, there must be edges between X and Z and between Y and Z owing to the reason given in Example 2. We cannot have any of the causal DAGs in Figure

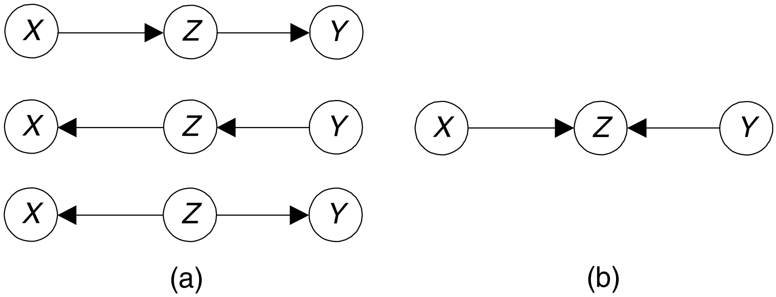


Figure 2.8: If the set of conditional independencies is $\{I(X,Y)\}$, we must have the causal DAG in (b).

2.8 (a). The reason is that the Markov condition, applied to these DAGs, entails $I(X,Y|Z)$, and this conditional independency is not present. So the Markov condition would not be satisfied. Furthermore, the Markov condition, applied to these DAGs, does not entail $I(X,Y)$. So the causal DAG must be the DAG in Figure 2.8 (b). We conclude that X and Y each cause Z.

Example 4 *Suppose* $\mathsf{V} = \{X,Y,Z\}$ *and our set of conditional independencies is*

$$\{I(X,Y|Z)\}.$$

Then owing to reasons similar to those given before, the only edges in the causal DAG must be between X and Z and between Y and Z. We cannot have the causal DAG in Figure 2.8 (b). The reason is that the Markov condition, applied to this DAG, entails $I(X,Y)$, and this conditional independency is not present. So the Markov condition would not be satisfied. So we must have one of the causal DAGs in Figure 2.8 (a).

We now state the following theorem, whose proof can be found in [11]. At this point your intuition should suspect that it is true.

Theorem 1 *If $(\mathbb{G},P)$ satisfies the faithfulness condition, then there is an edge between X and Y if and only if X and Y are not conditionally independent given any set of variables.*

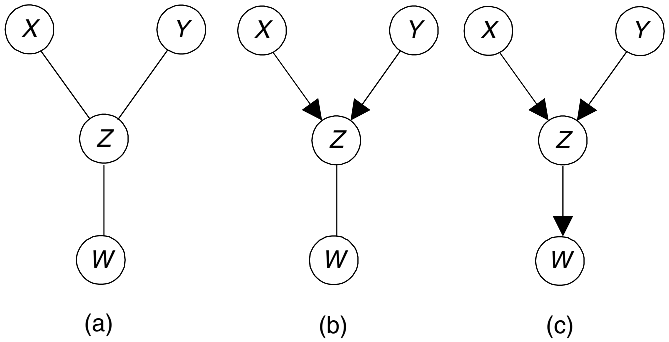


Figure 2.9: If the set of conditional independencies is $\{I(X,Y),\quad I(W,\{X,Y\}|Z)\}$, we must have the causal DAG in (c).

Example 5 *Suppose* $\mathsf{V} = \{X,Y,Z,W\}$ *and our set of conditional independencies is*

$$\{I(X,Y),\quad I(W,\{X,Y\}|Z)\}.$$

Owing to Theorem 1, the links (edges without regard for direction) must be as shown in Figure 2.9 (a). We must have the directed edges shown in Figure 2.9 (b) because we have $I(X,Y)$*. Therefore, we must also have the directed edge shown in Figure 2.9 (b) because we do not have* $I(W,X)$*. We conclude* X *and* Y *each cause* Z *and* Z *causes* W*.*

Example 6 *Suppose* $\mathsf{V} = \{X,Y,Z,W\}$ *and our set of conditional independencies is*

$$\{I(X,\{Y,W\}),\quad I(Y,\{X,Z\})\}.$$

Owing to Theorem 1, we must have the links shown in Figure 2.10 (a). Now if we have the chain $X \rightarrow Z \rightarrow W$, $X \leftarrow Z \leftarrow W$, *or* $X \leftarrow Z \rightarrow W$, *then we would not have* $I(X,W\}$, *which is an independency that can be readily deduced from* $I(X,\{Y,W\})$*. So we must have the chain* $X \rightarrow Z \leftarrow W$*. Similarly, we must have chain* $Y \rightarrow W \leftarrow Z$*. So our causal graph must be the one in Figure 2.10 (b). However,*

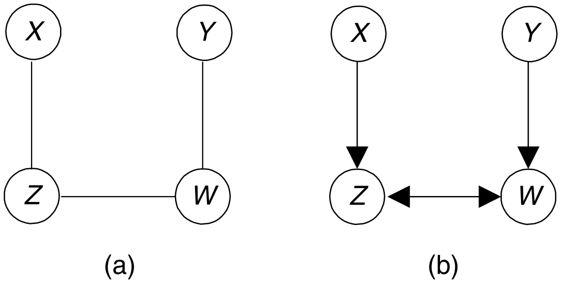


Figure 2.10: If the set of conditional independencies is $\{I(X,\{Y,W\}),\ I(Y,\{X,Z\})\}$, we must have the causal graph in (b).

this graph is not a DAG. The problem here is that this probability distribution does not admit a faithful DAG representation, which tells us we cannot make the causal faithfulness assumption. In the next subsection we will revisit this example while making only the causal embedded faithfulness assumption.

2.2.2 Assuming Only Causal Embedded Faithfulness

Previously, we mentioned that the most problematic assumption in the causal faithfulness assumption is that there must be no hidden common causes, and we eliminated that problem with the causal embedded faithfulness assumption. Let's see how much we can learn when making only this assumption.

Example 7 *In Example 3 we had* $\mathsf{V} = \{X, Y, Z\}$ *and the set of conditional independencies*

$$\{I(X,Y)\}.$$

Under the causal faithfulness assumption, we concluded that X and Y each cause Z. However, the probability distribution is embedded faithfully in all the DAG in Figure 2.11. So it could be that X causes Z or it could be that X and Z have a hidden common cause. The same holds for Y and Z.

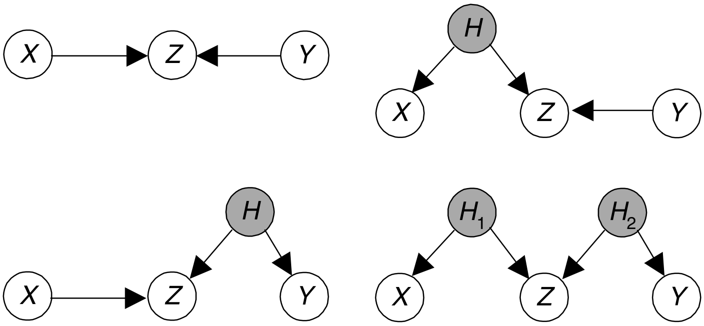


Figure 2.11: If we make the causal embedded faithfulness assumption and our set of conditional independencies is $\{I(X,Y)\}$, the causal relationships could be the ones in any of these DAGs.

While making only the more reasonable causal embedded faithfulness assumption, we were not able to learn any causal influences in the previous example. Can we ever learn a causal influence while making this assumption? The next example shows that we can.

Example 8 *In Example 5 we had* $\mathsf{V} = \{X, Y, Z, W\}$ *and the set of conditional independencies*

$$\{I(X,Y), \quad I(W,\{X,Y\}|Z)\}.$$

The probability distribution P *is embedded faithfully in the DAGs in Figure 2.12 (a) and (b). However, it is not embedded faithfully in the DAGs in Figure 2.12 (c) or (d). The reason is that these latter DAGs entail* $I(X,W)$*, and we do not have this conditional independency. That is, the Markov condition says* X *must be independent of its nondescendents conditional on its parents. Since* X *has no parents, this means* X *must simply be independent of its nondescendents, and* W *is one of its nondescendents. We conclude that* Z *causes* W*.*

Example 9 *In Example 6 we had* $\mathsf{V} = \{X, Y, Z, W\}$ *and the set of conditional independencies*

$$\{I(X,\{Y,W\}), \quad I(Y,\{X,Z\})\}.$$

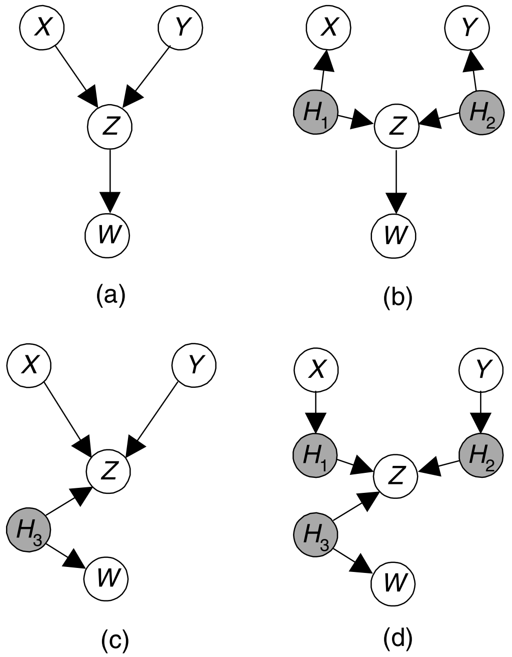


Figure 2.12: If our set of conditional independencies is $\{I(X,Y), \quad I(W,\{X,Y\}|Z)\}$, then P is embedded faithfully in the DAGs in (a) and (b) but not in the DAGs in (c) and (d).

Recall that we obtained the graph in Figure 2.13 (a) when we tried to find a DAG faithful to the probability distribution P. We concluded that P does not admit a faithful DAG representation. On the other hand, P is embedded faithfully in the DAGs in Figure 2.13 (b) and (c). We conclude that Z and W have a hidden common cause.

Example 10 *Suppose we have these variables:*

R: Parent's smoking history

A: Alcohol consumption

S: Smoking behavior

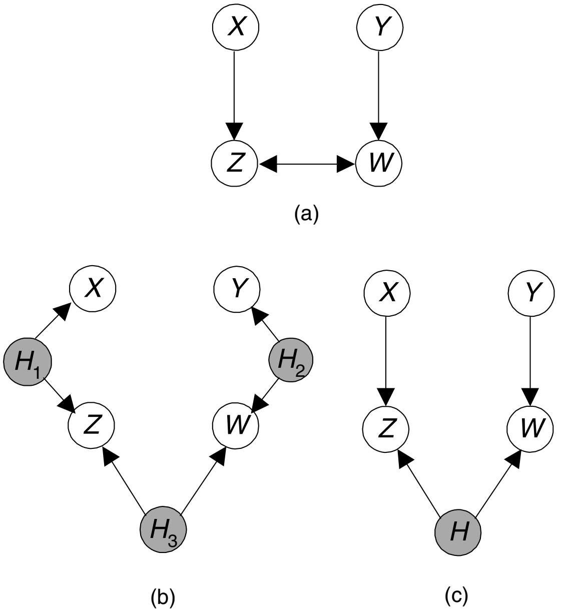


Figure 2.13: If our set of conditional independencies is $\{I(X, \{Y, W\}), \quad I(Y, \{X, Z\})$, we can conclude that Z and W have a hidden common cause.

L: Lung Cancer

Suppose further we learn the following conditional independencies from data:

$$\{I(R, A), \quad I(L, \{R, A\}|S)\}.$$

We conclude the causal relationships in Figure 2.14. In that figure the edges $R \rightarrow S$ and $A \rightarrow S$ mean that the first variable causes the second, or the two variables have a hidden common cause, or the first causes the second and they have a hidden common cause. The edge $S \rightarrowtail L$ means S causes L and they do not have a hidden common cause. We conclude smoking causes lung cancer. This example is only for the sake of illustration. We know of no data set indicating these conditional independencies.

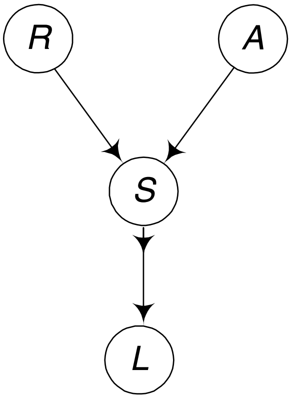


Figure 2.14: S has a causal influence on L.

Based on considerations such as those illustrated in the previous examples, Spirtes et al. [17] developed an algorithm that finds the causal DAG faithful to P from the conditional independencies in P when the causal faithfulness assumption is made. Meek [9] proved the correctness of the algorithm. Spirtes et al. [17] further developed an algorithm that learns causal influences from the conditional independencies in P when the causal embedded faithfulness assumption is made. They conjecture that it finds all possible causal influences among the observed variables. The algorithm is also described in detail in [11].

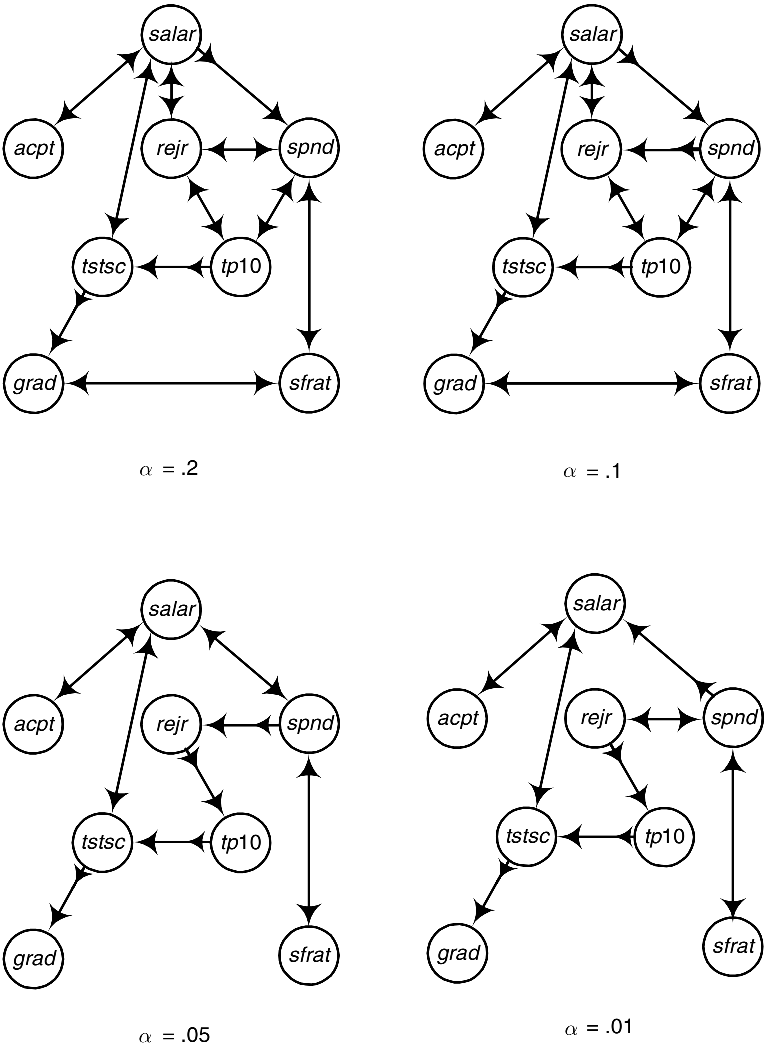
Next we show two studies that use the method just described to learn causal influences.

Example 11 *Using the data base collected by the U.S. News and World Record magazine for the purpose of college ranking, Druzdzel and Glymour [4] analyzed the influences that affect university student retention rate. By 'student retention rate' we mean the percent of entering freshmen who end up graduating from the university at which they initially matriculate. Low student retention rate is a major concern at many American universities as the mean retention rate over all American universities is only 55%.*

The data base provided by the U.S. News and World Record magazine contains records for 204 United States universities and colleges identified as major research institutions. Each record consists of over 100 variables. The data was collected separately for the years 1992 and 1993. Druzdzel and Glymour [4] selected the following eight variables as being most relevant to their study:

Variable	*What the Variable Represents*
grad	*Fraction of entering students graduating from the institution*
rejr	*Fraction of applicants who are not offered admission*
tstsc	*Average standardized score of incoming students*
tp10	*Fraction of incoming students in the top* 10% *in high school*
acpt	*Fraction of students accepting the institution's admission offer*
spnd	*Average educational and general expenses per student*
sfrat	*Student/faculty ratio*
salar	*Average faculty salary*

Druzdzel and Glymour [4] used Tetrad II [16] to learn causal influences from the data. Tetrad II uses the previously mentioned algorithm developed by Spirtes et al. [17] to learn causal structure from data. Tetrad II allows the user to enter a significance level. A significance level of α *means the probability of rejecting a conditional independency hypothesis, when it it is true, is* α. *Therefore, the smaller the value* α, *the less likely we are to reject a conditional independency, and therefore the sparser our resultant graph. Figure 2.15 shows the graphs, which Druzdzel and Glymour [4] learned from U.S. News and World Record's 1992 data base using significance levels of* .2, .1, .05, *and* .01. *In those graphs, an edge* $X \rightarrow Y$ *indicates either* X *has a causal influence on* Y, *or* X *and* Y *have a hidden common cause or both; an edge* $X \leftrightarrow Y$ *indicates* X *and* Y *have a hidden common cause; and an edge* $X \rightarrowtail Y$ *indicates* X *has a causal influence on* Y *and they do not have a hidden common cause.*



α = .2

α = .1

α = .05

α = .01

Figure 2.15: The graphs Tetrad II learned from *U.S. News and World Record*'s 1992 data base.

Although different graphs were obtained at different levels of significance, all the graphs in Figure 2.15 show that average standardized test score (tstsc) has a direct causal influence on graduation rate (grad), and no other variable has a direct causal influence on grad. The results for the 1993 data base were not as overwhelming, but they too indicated tstsc to be the only direct causal influence of grad.

To test whether the causal structure may be different for top research universities, Druzdzel and Glymour [4] repeated the study using only the top 50 universities according to the ranking of U.S. News and World Report. The results were similar to those for the complete data base.

These result indicate that, although factors such as spending per student and faculty salary may have an influence on graduation rates, they do this only indirectly by affecting the standardized test scores of matriculating students. If the results correctly model reality, retention rates can be improved by bringing in students with higher test scores in any way whatsoever. Indeed, in 1994 Carnegie Mellon changed its financial aid policies to assign a portion of its scholarship fund on the basis of academic merit. Druzdzel and Glymour [4] note that this resulted in an increase in the average test scores of matriculating freshman classes and an increase in freshman retention.

Before closing, we note that the notion that average test score has a causal influence on graduation rate does not fit into common notions of causation such as the one concerning manipulation. For example, if we manipulated a university's average test score by accessing the testing agency's database and changing the scores of the university's students to much higher values, we would not expect the university's graduation rate to increase. Rather this study indicates that test score is a near perfect indicator of some other variable, which we can call 'graduation potential'.

The next example, taken from [12] illustrates problems one can encounter when inferring causation from passive data.

Example 12 *Scarville et al. [15] provide a data base obtained from a survey in 1996 of experiences of racial harassment and discrimination of military personnel in the United States Armed Forces. Surveys were* *distributed to 73,496 members of the U.S. Army, Navy, Marine Corps, Air Force and Coast Guard. The survey sample was selected using a nonproportional stratified random sample in order to ensure adequate representation of all subgroups. Usable surveys were received from 39,855 service members (54%). The survey consisted of 81 questions related to experiences of racial harassment and discrimination and job attitudes. Respondents were asked to report incidents that had occurred during the previous 12 months. The questionnaire asked participants to indicate the occurrence of 57 different types of racial/ethnic harassment or discrimination. Incidents ranged from telling offensive jokes to physical violence, and included harassment by military personnel as well as the surrounding community. Harassment experienced by family members was also included.*

Neapolitan and Morris [12] used Tetrad III in an attempt to learn causal influences from the data base. For their analysis, 9640 records (13%) were selected which had no missing data on the variables of interest. The analysis was initially based on eight variables. Similar to the situation concerning university retention rates, they found one causal relationship to be present regardless of the significance level. That is, they found that whether the individual held the military responsible for the racial incident had a direct causal influence on the individual's race. Since this result made no sense, they investigated which variables were involved in Tetrad III learning this causal influence. The five variables involved are the following:

Variable	*What the Variable Represents*
race	*Respondent's race/ethnicity*
yos	*Respondent's years of military service*
inc	*Did respondent experience a racial incident?*
rept	*Was incident reported to military personnel?*
resp	*Did respondent hold military responsible for incident?*

The variable race consisted of five categories: White, Black, Hispanic, Asian or Pacific Islander, and Native American or Alaskan Native.

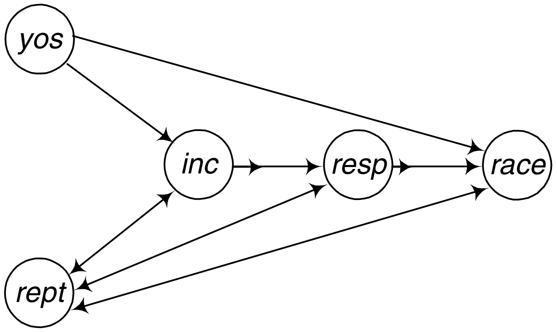


Figure 2.16: The graph Tetrad III learned from the racial harassment survey at the .01 significance level.

Respondents who reported Hispanic ethnicity were classified as Hispanic, regardless of race. Respondents were classified based on self-identification at the time of the survey. Missing data were replaced with data from administrative records. The variable yos was classified into four categories: 6 years or less, 7-11 years, 12-19 years, and 20 years or more. The variable inc was coded dichotomously to indicate whether any type of harassment was reported on the survey. The variable rept indicates responses to a single question concerning whether the incident was reported to military and/or civilian authorities. This variable was coded 1 *if an incident had been reported to military officials. It was coded* 0 *if an individual experienced no incident, did not report the incident, or only reported the incident to civilian officials. The variable resp indicates responses to a single question concerning whether the respondent believed the military to be responsible for an incident of harassment. This variable was coded* 1 *if the respondent indicated that the military was responsible for some or all of a reported incident. If the respondent indicated no incident, unknown responsibility, or that the military was not responsible, the variable was coded* 0*.*

They reran the experiment using only these five variables, and again at all levels of significance, they found that resp had a direct *causal influence on race. In all cases, this causal influence was learned because rept and yos were found to be probabilistically independent, and there was no edge between race and inc. That is, the causal connection between race and inc is mediated by other variables. Figure 2.16 shows the graph obtained at the* .01 *significance level. The edges yos → inc and rept → inc are directed towards inc because yos and rept were found to be independent. The edge yos → inc resulted in the edge inc ↣ resp being directed the way it was, which in turn resulted in resp ↣ race being directed the way it was. If there had been an edge between inc and race, the edge between resp and race would not have been directed.*

It seems suspicious that no direct causal connection between race and inc was found. Recall, however, that these are the probabilistic relationships among the responses; they are not necessarily the probabilistic relationships among the actual events. There is a problem with using responses on surveys to represent occurrences in nature because subjects may not respond accurately. This is called response bias. Let's assume race is recorded accurately. The actual causal relationship between race, inc, and says_inc may be as shown in Figure 2.17. By inc we now mean whether there really was an incident, and by says_inc we mean the survey response. It could be that races, which experienced higher rates of harassment, were less likely to report the incident, and the causal influence of race on says_inc through inc was negated by the direct influence of race on inc. The previous conjecture is substantiated by another study. Stangor et al. [18] examined the willingness of people to attribute a negative outcome to discrimination when there was evidence that the outcome might be influenced by bias. They found that minority members were more likely to attribute the outcome to discrimination when responses were recorded privately, but less likely to report discrimination when they had to express their opinion publicly and there was a member of the non-minority group present. This suggests that while minorities are more likely to perceive the situation as due to discrimination, they are less likely to report it publicly. Although the survey of military personnel was intended to be confidential, minority members in the military may have felt uncomfortable reporting incidents of discrimination.

Tetrad III allows the user to enter a temporal ordering. So Neapoli-

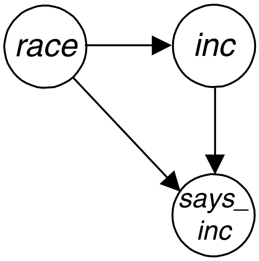


Figure 2.17: Possible causal relationships among race, incidence of harassment, and saying there is an incident of harassment.

tan and Morris [12] could have put race first in such an ordering to avoid it being an effect of another variable. However, one should do this with caution. The fact that the data strongly supports that race is an effect indicates there is something wrong with the data, which means one should be dubious of drawing any conclusions from the data. In the present example, Tetrad III actually informed them that they could not draw causal conclusions from the data when they made race a root. That is, when they made race a root, Tetrad III concluded there is no consistent orientation of the edge between race and resp.

2.2.3 Assuming Causal Embedded Faithfulness with Selection Bias

In the previous subsection we deduced causal influences assuming selection bias is not present. Here we relax that assumption. If we assume the probability distribution P of the observed variables is embedded faithfully in a causal DAG containing the variables, but that possibly selection bias is present when we sample, we say we are making the **causal embedded faithfulness assumption with selection bias**.

Before showing an example of causal learning under this assumption, let's discuss selection bias further. Recall in Section 2.1.2 we mentioned that the pharmaceutical company Merck noticed that its

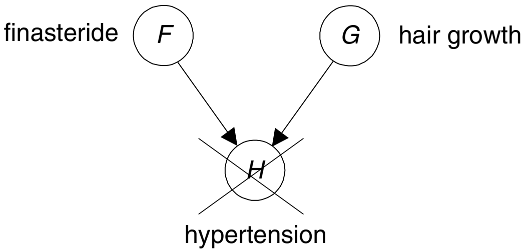


Figure 2.18: The instantiation of H creates a dependency between F and G.

drug finasteride appeared to cause hair regrowth. Now suppose finasteride (F) and apprehension about lack of hair regrowth (G) are both causes of hypertension (H), and Merck happened to be observing individuals who had hypertension. We say a node is **instantiated** when we know its value for the entity currently being modeled. So we are saying H is instantiated to the same value for all entities in the population we are observing. This situation is depicted in Figure 2.18, where the cross through H means the variable is instantiated. Ordinarily, the instantiation of a common effect creates a dependency between its causes because each cause explains away the occurrence of the effect, thereby making the other cause less likely. Psychologists call this **discounting**. So, if this were the case, discounting would explain the correlation between F and G. As mentioned previously, this type of dependency is called **selection bias**. This example is only for the sake of illustration. There is no evidence that finasteride causes hypertension, and it apparently does cause hair regrowth by lowering DHT levels.

The next example shows that we can learn causal influences even when selection bias may be present.

Example 13 *Suppose we have the same variables and distribution as in Examples 5 and 8, and now we assume selection bias may be present. Recall* $\mathsf{V} = \{X, Y, Z, W\}$ *and our set of conditional independencies in the observed probability distribution* P *(the one obtained*

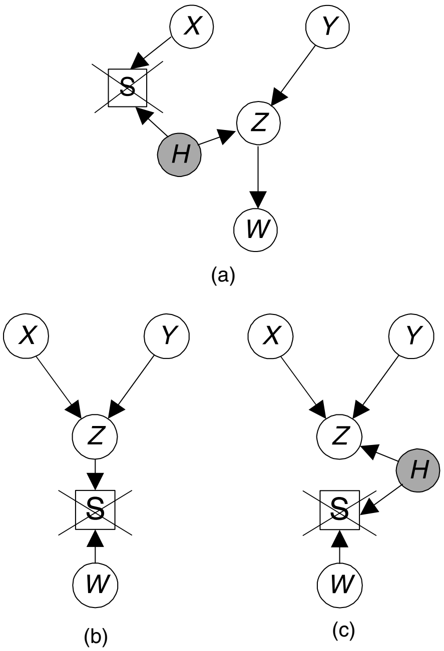


Figure 2.19: If the observed set of conditional independencies is $\{I(X,Y),\quad I(W,\{X,Y\}|Z)\}$, the actual probability distribution could only be embedded faithfully in the DAG in (a).

when selection bias may be present) was

$$\{I(X,Y),\quad I(W,\{X,Y\}|Z)\}.$$

In Example 8, when making the assumption of causal embedded faithfulness, we concluded that Z causes W. Next we show that we can conclude this even when we only make the causal embedded faithfulness assumption with selection bias. The actual probability distribution P' (the one obtained when S is not instantiated and so there is no selection bias) could be embedded faithfully in the causal DAG in Figure 2.19 (a), but it could not be embedded faithfully in the casual DAG in Figure 2.19 (b) or (c). The reason is that, owing to the instantiation of S, the DAG in Figure 2.19 (b) does not entail $I(X,Y)$ in *the observed distribution P, and the DAG in Figure 2.19 (c) does not entail $I(W,\{X,Y\}|Z)$ in the observed distribution P. So we can still conclude that Z has a causal influence on W.*

2.3 Learning Causation From Data on Two Variables

The previous section discussed learning causal structure from the conditional independencies in the joint probability distribution of the variables. These conditional independencies are obtained from the data in a random sample. Another way to learn causal structure is to develop a scoring function *score* (called a **scoring criterion**) that assigns a value $score(\mathsf{data},\mathbb{G})$ to each causal DAG under consideration based directly on the data. Ideally we want a scoring criterion to be consistent. First we discuss consistency and some other preliminary concepts. Then we apply the method to causal learning.

2.3.1 Preliminary Concepts

A **DAG model** consists of a DAG $\mathbb{G}=(\mathsf{V},\mathsf{E})$, where V is a set of random variables, and a parameter set F whose members determine conditional probability distributions for the DAG, such that for every permissible assignment of values to the members of F, a joint probability distribution of V is given by the product of these conditional distributions and this joint probability distribution satisfies the Markov condition with the DAG. For simplicity, we ordinarily denote a DAG model using only $\mathbb{G}$ (i.e. we do not show F.) Probability distribution P is **included** in model $\mathbb{G}$ if there is some assignment of values to the parameters in F that yields the probability distribution.

Example 14 *DAG models appear in Figures 2.20 (a) and (b). The values of X, Y, and Z are $x1$, $x2$, $y1$, $y2$, $y3$, $z1$, and $z2$. The probability distribution contained in the causal network in Figure 2.20 (c) is included in both models, whereas the one in the causal network in Figure 2.20 (d) is included only in the model in Figure 2.20 (b). The reason this latter probability distribution is not included in the model*

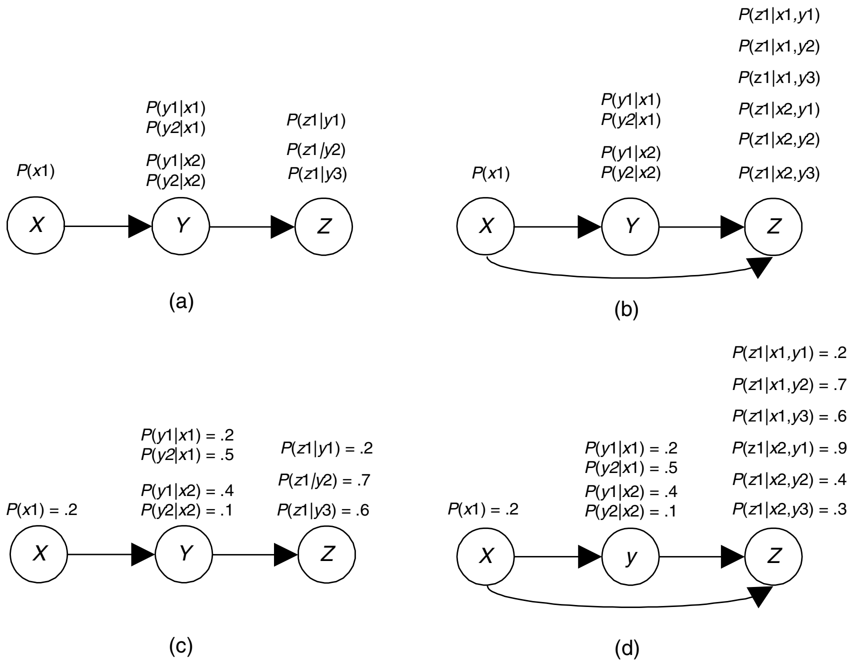


Figure 2.20: DAG models appear in (a) and (b). The probability distribution in the causal network in (c) is included in both models, whereas the one in (d) is included only in the model in (b).

in Figures 2.20 (a) is that any probability distribution included in that model would need to have $I(X, Z|Y)$, and this probability distribution does not have that conditional independency.

The **dimension** of a DAG model is the number of parameters in the model. The dimension of the DAG model in Figures 2.20 (a) is 8, while the dimension of the one in Figures 2.20 (b) is 11.

We now have the following definition concerning scoring criteria:

Definition 1 *Let data_n be a set of values (data) of a set of n mutually independent random vectors, each with probability distribution P, and let P_n be the probability function determined by the joint distribution* *of the n random vectors. Furthermore, let score be a scoring criterion over some set of DAG models for the random variables that constitute each vector. We say score is* ***consistent*** *for the set of models if the following two properties hold:*

1. *If* $\mathbb{G}_1$ *includes* P *and* $\mathbb{G}_2$ *does not, then*

$$\lim_{n\to\infty} P_n\left(score(\mathsf{data}_n, \mathbb{G}_1) > score(\mathsf{data}_n, \mathbb{G}_2)\right) = 1.$$

2. *If* $\mathbb{G}_1$ *and* $\mathbb{G}_2$ *both include* P *and* $\mathbb{G}_1$ *has smaller dimension than* $\mathbb{G}_2$*, then*

$$\lim_{n\to\infty} P_n\left(score(\mathsf{data}_n, \mathbb{G}_1) > score(\mathsf{data}_n, \mathbb{G}_2)\right) = 1.$$

We call P the **generative distribution**. The limit, as the size of the data set approaches infinity, of the probability that a consistent scoring criterion chooses a smallest model that includes P is 1.

The **Bayesian scoring criterion,** which is the

$$P(\mathsf{data}|\mathbb{G})$$

(i.e. it is the probability of the data given the DAG), is a consistent scoring criterion for DAG models whose parameters are discrete. This criterion and its consistency are discussed in detail in [11]. Presently, we just show a few simple examples illustrating the result of applying it.

Example 15 *Suppose* $\mathsf{V} = \{X, Y\}$*, both variables are binary, and the set of conditional independencies in probability distribution* P *is*

$$\{I(X,Y)\}.$$

Let the values of X *be* $x1$ *and* $x2$ *and the values of* Y *be* $y1$ *and* $y2$*. Possible DAG models are shown in Figure 2.21. Both models include* P*. When the data set is large, the model in (a) should be chosen by the Bayesian scoring criterion because it has smaller dimension.*

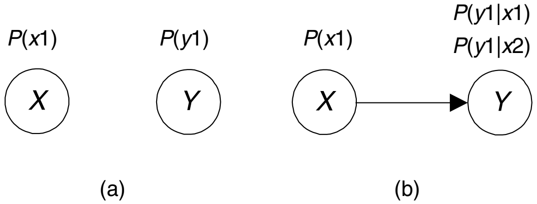


Figure 2.21: Two DAG models.

Example 16 *Suppose* $\mathsf{V} = \{X, Y\}$*, both variables are binary, and the set of conditional independencies in probability distribution* P *is the empty set*

$$\varnothing.$$

Possible DAG models are shown in Figure 2.21. When the data set is large, the model in (b) should be chosen by the Bayesian scoring criterion because it is the only one that includes P*.*

A **hidden variable DAG model** is a DAG model augmented with hidden variables. Figure 2.22 (b) shows a hidden variable DAG model. The variables that are not hidden are called **observables**. It has not been proven whether, in general, the Bayesian scoring criterion is consistent when hidden variable DAG models are also considered. However, Rusakov and Geiger [14] proved it is consistent in the case of **naive hidden variable DAG models**, which are models such that there is a single hidden variable H, all observables are children of H, and there are no edges between any observables.

Example 17 *Suppose* $\mathsf{V} = \{X, Y\}$*, both variables have three possible values, and the set of conditional independencies in probability distribution* P *is the empty set*

$$\varnothing.$$

Let $\mathbb{G}$ *be the DAG model in Figure 2.22 (a) and* $\mathbb{G}_H$ *be the naive hidden variable DAG model in Figure 2.22 (b). Although* $\mathbb{G}_H$ *appears larger, it is actually smaller because it has fewer effective parameters.*

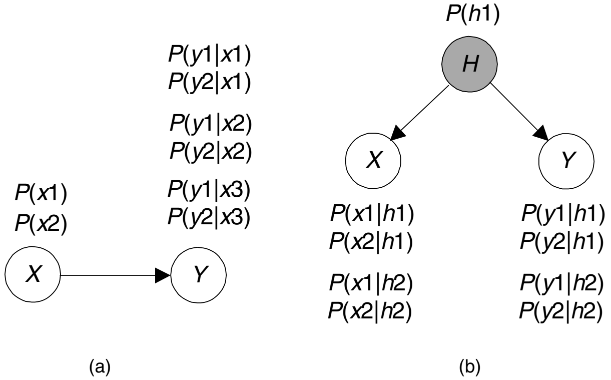


Figure 2.22: A DAG model and a hidden variable DAG model.

This is discussed rigorously in [11]. The following is an intuitive explanation. Since $\mathbb{G}$ includes all joint probability distributions of these variables, $\mathbb{G}$ includes every distribution that $\mathbb{G}_H$ includes. However, $\mathbb{G}_H$ only includes distributions that can be represented by urn problems in which there is a division of the objects into two sets such that X and Y are independent in each set. For example, $\mathbb{G}_H$ includes the joint distribution of the value (1, 2, or 3) and shape (circle, square, or arrow) of the objects Figure 2.23 (a) because value and shape are independent given either the set of black objects or the set of white objects. However, $\mathbb{G}_H$ does not include the joint distribution of the value and shape of the objects in Figure 2.23 (b) because there is no division of the objects into two sets that value and shape are independent in each set.

As discussed in [11], in the space consisting of all possible assignments of values to the parameters of model $\mathbb{G}$ in Figure 2.22 (a), the subset, whose probability distributions are included in hidden variable model $\mathbb{G}_H$ in Figure 2.22 (b), has Lebesgue measure zero. This means that if we assign arbitrary values to the parameters in $\mathbb{G}$, we can be almost certain the resultant probability distribution is not included in $\mathbb{G}_H$.

Consider now the following experiment:

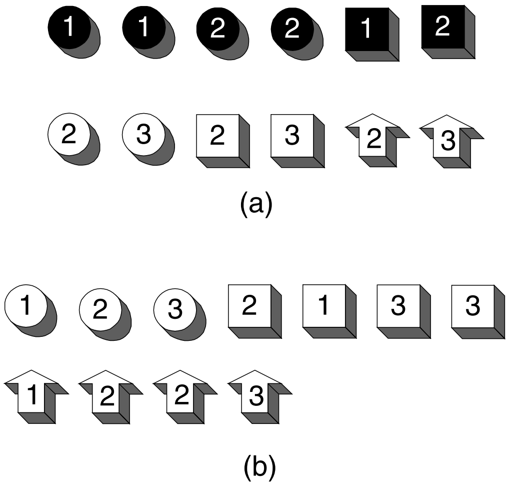


Figure 2.23: Value and shape are independent given color in (a). There is no division of the objects in (b) into two sets which renders value and shape independent.

1. Randomly choose either $\mathbb{G}$ or $\mathbb{G}_H$.

2. Randomly assign parameter values to the model chosen.

3. Generate a large amount of data.

4. Score the DAGs based on the data using the Bayesian scoring criterion.

When $\mathbb{G}$ is chosen, almost certainly the probability distribution will not be included in $\mathbb{G}_H$. So with high probability $\mathbb{G}$ will score higher. When $\mathbb{G}_H$ is chosen, with high probability $\mathbb{G}_H$ will score higher because it has smaller dimension than $\mathbb{G}$. So from the data alone we can become very confident as to which DAG model was randomly chosen.

2.3.2 Application to Causal Learning

Next we show how the theory just developed can be applied to causal learning. Suppose some large population is distributed according to the data in the following table:

Case	Sex	Height (inches)	Wage (\$)
1	*female*	64	30,000
2	*female*	64	30,000
3	*female*	64	40,000
4	*female*	64	40,000
5	*female*	68	30,000
6	*female*	68	40,000
7	*male*	64	40,000
8	*male*	64	50,000
9	*male*	68	40,000
10	*male*	68	50,000
11	*male*	70	40,000
12	*male*	70	50,000

The random variables Sex, $Height$, $Wage$ have the same joint probability distribution as the random variables $Color$, $Shape$, and $Value$ in Figure 2.23 (a) when we make the following associations:

$$black/female, \quad white/male, \quad circle/64, \quad square/68,$$

$$arrow/70, \quad 1/30,000 \quad 2/40,000, \quad 3/50,000.$$

Suppose now we only observe and collect data on height and wage. Sex is then a hidden variable in the sense that it renders the observed variables independent. If we only looked for correlation, we would find height and wage are correlated and perhaps conclude height has a causal effect on wage. However, if we score the models $\mathbb{G}$ and $\mathbb{G}_H$ in Figure 2.22, $\mathbb{G}_H$ will most probably win because it has smaller dimension than $\mathbb{G}$. We can then conclude that possibly there is a hidden common cause.

We said 'possibly' because there are a number of caveats when concluding causation from data on two variables. They are as follows:

1. The hidden variable DAG models $X \rightarrow H \rightarrow Y$ and $X \leftarrow H \leftarrow Y$ have the same score as $X \leftarrow H \rightarrow Y$. So we may have a hidden intermediate cause instead of a hidden common cause.

2. In real applications features like height and wage are continuous. We create discrete values by imposing cutoff points. With different cutoff points we may not create a division that renders wage and height independent.

3. Similar caution must be used if the DAG model wins and we want to conclude X causes Y or Y causes X. That is,

(**a**) Different cutoff points may result in a division of the objects that renders them independent, which means the hidden variable model would win.

(**b**) If there is a hidden common cause, it may be modeled better with a hidden variable that has a larger space. Clearly, if we increase the space size of H sufficiently the hidden variable model will have the same size as the DAG model.

(**c**) Selection bias may be present. For example, if X causes Z and Y causes Z, and we are observing a population in which all members have $Z = z'$, then the observed distribution is $P(x, y|z') = P(y|x, z')P(x|z')$, which means the observed distribution is included in the DAG $X \rightarrow Y$.

We conclude that data on two variables can only give us an indication as to what may be going on causally. For example, in the example involving sex, height, and wage, the data can inform us that it seems there may be a binary hidden common cause. Given this indication, we can then investigate the situation further by searching for one.

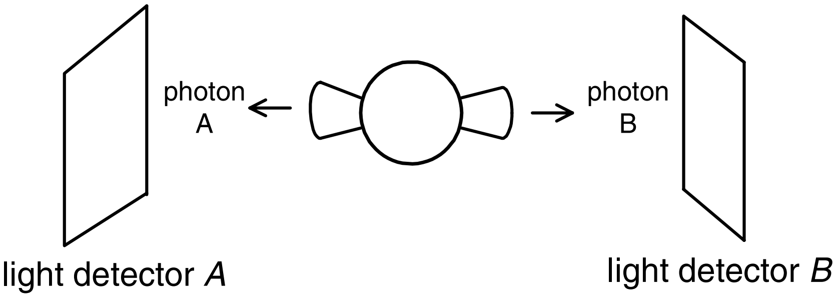


Figure 2.24: Dual photons are emitted in opposite directions.

2.3.3 Application to Quantum Mechanics

The theory of quantum mechanics entails that not all physical observables can be simultaneously known with unlimited precision, even in principle. For example, you can place a subatomic particle so its position is well-defined, but then you cannot determine its momentum. If the momentum is well-defined, you cannot determine its position. It's not just a human measurement problem, but rather that precise values cannot simultaneously be assigned to both momentum and position in the mathematics of quantum mechanics. The assignment of a precise value to momentum determines a unique probability density function of position and vice versa.

The EPR Paradox

According to Einstein, Podolsky, and Rosen [5], a 'complete theory' would be able to represent the simultaneous existence of any properties that do simultaneously exist. They set out to show that properties like momentum and position do simultaneously exist and that therefore quantum mechanics is an incomplete theory. To that end, they developed a thought experiment, called the **EPR Paradox**, similar to the one we describe next. Their thought experiment involved position and momentum. However, since the actual experiments that were eventually done involved spin, we describe this latter experiment.

Figure 2.24 depicts a situation in which a pair of photons is emitted, and measurements are made an arbitrary distance away, which could be many light years. Each photon has a spin along each of the x-axis, y-axis, and z-axis. If photon A's spin on the x-axis is positive, photon B's must be negative. So by measuring photon A's spin on any axis, we can learn Photon B's spin on that axis. Experiments, in which we measured, for example, both spins on the x-axis, have substantiated that they are opposite of each other. Like position and momentum, the three spins of a photon cannot 'simultaneously exist' in quantum mechanics.

Einstein et al. [5] made the 'locality assumption', which says no instantaneous change in photon B can occur owing to a measurement at photon A. So the value of the spin of photon B must have existed before the measurement at A took place. Since we can do this experiment for each of the spin directions, this is true for all three spins. So the values of the three spins must simultaneously exist. We conclude that the spins must have been determined at the time the split was made, and that the conditions at this time constitute a hidden variable that renders the spins of photon A and B independent. We conclude quantum mechanics is incomplete.

Bell's Inequality

In 1964 John Bell [2] showed that if there is a hidden variable, then the probability distribution of the spins must satisfy Bell's Inequality. In our terminology, this means the hidden variable DAG model only includes probability distributions that satisfy Bell's Inequality. However, a number of experiments have shown that Bell's Inequality is not satisfied [1]. We discuss Bell's Inequality, and these results and their ramifications next.

Theorem 2 *(Bell's Inequality) Let*

1. X, Y, and Z be any three properties of entities in some population.

2. $\#(X, Y)$ denote the number of entities with properties X and Y.

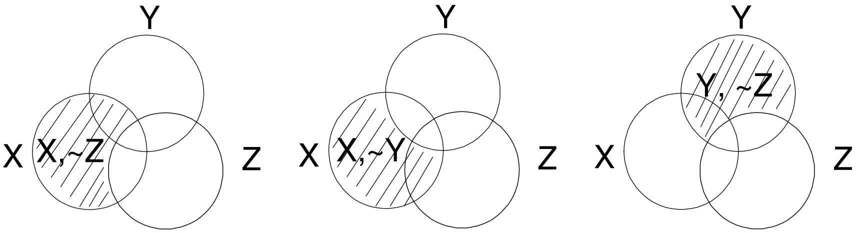


Figure 2.25: A proof of Bell's Inequality.

Then

$$\#(X,^{\sim} Z) \leq \#(X,^{\sim} Y) + \#(Y,^{\sim} Z),$$

where '∼' is the Not operator.

Proof. *The proof follows from the Venn diagram in Figure 2.25.* ■

Now let X_A be a variable whose values are the possible spins of photon A in the x-axis. These values are $x_A \uparrow$ and $x_A \downarrow$. Let similar definitions hold for Y, Z, and B. Then if we emit many photons, according to Bell's Inequality,

$$\#(x_A \uparrow, z_A \downarrow) \leq \#(x_A \uparrow, y_A \downarrow) + \#(y_A \uparrow, z_A \downarrow).$$

Since B's directions are always opposite to A's directions, this means

$$\#(x_A \uparrow, z_B \uparrow) \leq \#(x_A \uparrow, y_B \uparrow) + \#(y_A \uparrow, z_B \uparrow)$$

These values we can actually measure. That is, we do one measurement at each detector. We cannot obtain all measurements simultaneously. However, we can repeat the experiment many times, each time taking one pair of measurements. This has been done quite a few times, and the inequality was found not to hold [1].

Next suppose, as depicted in Figure 2.26, we have a hidden variable generating six spin values according to some probability distribution of the six spin variables, and this distribution has the marginal distributions

$$P(x_A, z_B), \quad P(x_A, y_B), \quad P(y_A, z_B).$$

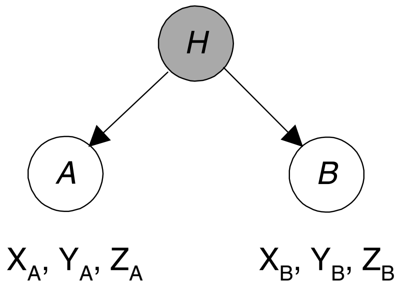


Figure 2.26: A hidden variable is generating values of the six spin variables.

Then if we have n repetitions of each measurement with n large,

$$\begin{aligned}
\#(x_A \uparrow, z_B \uparrow) &\approx P(x_A \uparrow, z_B \uparrow) \times \ n \\
\#(x_A \uparrow, y_B \uparrow) &\approx P(x_A \uparrow, y_B \uparrow) \times \ n \\
\#(y_A \uparrow, z_B \uparrow) &\approx P(y_A \uparrow, z_B \uparrow) \times \ n
\end{aligned}$$

and these numbers would have to obey Bell's Inequality. Since, as noted above, the observed numbers do not obey Bell's Inequality, we conclude the observed distributions are not marginals of distributions that are included in the hidden variable DAG model in Figure 2.26. Note that the observed distributions cannot be marginals of any distribution of the six variables. So we also cannot suppose that, at the time of emission, a probability distribution of the six variables is being generated by photon A obtaining three spin values and these values then causing photon B to have three values. A similar statement holds for selection bias.

Note that it may seem that we have 'cheated' quantum mechanics. That is, when we measure x_A and z_B, we learn the values of x_A and z_A (since z_A must be in the opposite direction of z_B). But quantum mechanics says we can't know both those values simultaneously. The following explanation clears this matter up:

> This was one of Schrödinger's first reactions to EPR in 1935. The problem is that when you measure x_A on one

> side and z_B on the other, there is no reason to think that the conservation law continues to hold, since you will have disturbed both systems. There is a little proof that it can't hold due to Asher Peres, buried in a little paper of mine.
> - Arthur Fine (private correspondence).

That is, when we measure x_A we disturb photon A; so its z_A value need no longer be the inverse of z_B.

In spite of this disturbance, the hidden variable model is still refuted. That is, if values of all six spin variables had been generated according to some probability distribution before the measurements, then the value of z_B would have been the inverse of the value of z_A right before the measurements. This means the (accurate) measured value of z_B would be the inverse of the value of z_A before the measurements.

So experimental results show that a hidden variable is not possible. Given this, what could explain the correlation between the measurements at the two detectors? Some explanations follow:

1. The measurement at detector A is 'causing' the values of the spins at detector B from many miles away and faster than the speed of light.

2. If you don't insist the spins of the two photons are separates variables, then you don't have any problem with superluminal causation or the violation of independence conditions. That is, although we thought we had two separate properties, we don't. Not only can't we figure out a way to vary them independently, but we have good theoretical reason for thinking we can't. In other words, we have a single variable (whose value are the three spin directions). The odd thing is that this property is nonlocal. That is, the measurement applies to a small region over here, and a small region over there, and nowhere in between - Charles Twardy.

3. We measure values for the left photon, and then the particle travels backwards in time to the source, collides with the source, which then emits the right photon. - Phil Dowe.

If we accept a simple manipulation definition of causation, there is no problem with the first and second explanations above. The two explanations are actually about the same because, in general, in a causal network a deterministic relationship between two variables can be modeled as a single variable. These explanations become difficult only when we insist that causation involves contact, material transfer, etc. Because our macro experiences, which gave rise to the concept of causality, have all involved physical contact, it seems our intuition requires it. However, if we had many experiences with manipulation similar to the ones at the quantum mechanical level, we may have different intuition. Finally, there is an epistemological explanation. Just as humans have the capability to view and therefore model more of reality than the amoeba, perhaps some more advanced creature could better view and model this result and thereby have a better intuition for it than us.

References

1. Aspect A, Grangier P, and Roger G (1982) Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedanken Experiment: A New Violation of Bell's Inequality, *Physical Review Letters*, 49 #2.

2. Bell JS (1964) On the Einstein-Podolsky-Rosen paradox, *Physics*, 1.

3. Christensen R (1990) *Log-Linear Models*, Springer-Verlag, New York.

4. Druzdzel MJ and Glymour C (1999) Causal Inference from Databases: Why Universities Lose Students, in Glymour C, and Cooper GF (eds.) *Computation, Causation, and Discovery*, AAAI Press, Menlo Park, California.

5. Einstein A, Podolsky B, and Rosen N (1935) Can Quantum-Mechanical Description of Physical Reality be Consdiered Complete, *Physical Review*, 47.

6. Heckerman D, Meek C, and Cooper G (1999) A Bayesian Approach to Causal Discovery, in Glymour C, and Cooper GF (eds.) *Computation, Causation, and Discovery*, AAAI Press, Menlo Park, California.

7. Lugg JA, Raifer J, and González CNF (1995) Dehydrotestosterone is the Active Androgen in the Maintenance of Nitric Oxide-Mediated Penile Erection in the Rat, *Endocrinology*, 136(4).

8. McClennan KJ and Markham A (1999) Finasteride: A review of its Use in Male Pattern Baldness, *Drugs*, 57(1).

9. Meek C (1995) Causal Influence and Causal Explanation with Background Knowledge, in Besnard P and Hanks S (eds.) *Uncertainty in Artificial Intelligence; Proceedings of the Eleventh Conference*, Morgan Kaufmann, San Mateo, California.

10. Neapolitan RE (1990) *Probabilistic Reasoning in Expert Systems*, Wiley, New York.

11. Neapolitan RE (2003) *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, New Jersey.

12. Neapolitan RE and Morris S (2004) Probabilistic Modeling Using Bayesian Networks, in Kaplan D (ed.) *Handbook of Quantitative Methodology in the Social Sciences*, Sage, Thousand Oaks, California.

13. Pearl J (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, United Kingdom.

14. Rusakov D, and Geiger D (2002) Bayesian Model Selection for Naive Bayes Models, in Darwiche A, and Friedman N (eds.) *Uncertainty in Artificial Intelligence; Proceedings of the Eighteenth Conference*, Morgan Kaufmann, San Mateo, California.

15. Scarville J, Button SB, Edwards JE, Lancaster AR, and Elig TW (1996) Armed Forces 1996 Equal Opportunity Survey, Defense Manpower Data Center, Arlington, VA. DMDC Report No. 97-0279.

16. Scheines R, Spirtes P, Glymour C, and Meek C (1994) *Tetrad II: User Manual*, Lawrence Erlbaum, Hillsdale, New Jersery.

17. Spirtes P, Glymour C, and Scheines E (1993, 2000) *Causation, Prediction, and Search*, Springer-Verlag, New York; 2nd ed.: MIT Press, Cambridge, Massachusetts.

18. Stangor C, Swim JK, Van Allen KL, and Sechrist GB (2002) Reporting Discrimination in Public and Private Contexts, *Journal of Personality and Social Psychology*, 82.