15:14 Friday 1st April, 2016

Chapter 24

Graphical Causal Models

[[TODO: discuss latent variables and measurement either here or in the graphical models chapter]]

24.1 Causation and Counterfactuals

Take a piece of cotton, say an old rag. Apply flame to it; the cotton burns. We say the fire *caused* the cotton to burn. The flame is certainly *correlated* with the cotton burning, but, as we all know, correlation is not causation (Figure 24.1). Perhaps every time we set rags on fire we handle them with heavy protective gloves; the gloves don't make the cotton burn, but the statistical dependence is strong. So what is causation?

We do not have to settle 2500 years (or more) of argument among philosophers and scientists. For our purposes, it's enough to realize that the concept has a **counterfactual** component: if, contrary to fact, the flame had not been applied to the rag, then the rag would not have burned[1]. On the other hand, the fire makes the cotton burn whether we are wearing protective gloves or not.

To say it a somewhat different way, the distributions we observe in the world are the outcome of complicated stochastic processes. The mechanisms which set the value of one variable inter-lock with those which set other variables. When we make a probabilistic prediction by conditioning — whether we predict $\mathbb{E}[Y \mid X = x]$ or $\Pr(Y \mid X = x)$ or something more complicated — we are just filtering the output of those mechanisms, picking out the cases where they happen to have set X to the value x, and looking at what goes along with that.

When we make a *causal* prediction, we want to know what would happen if the usual mechanisms controlling X were suspended and it was *set* to x. How would this change propagate to the other variables? What distribution would result for Y? This is often, perhaps even usually, what people really want to know from a data analysis, and they settle for statistical prediction either because they think it *is* causal prediction, or for lack of a better alternative.

Causal inference is the undertaking of trying to answer causal questions from empirical data. Its fundamental difficulty is that we are trying to derive counterfactual conclusions with only factual premises. As a matter of habit, we come to

[1] If you immediately start thinking about quibbles, like "What if we hadn't applied the flame, but the rag was struck by lightning?", then you may have what it takes to be a philosopher.

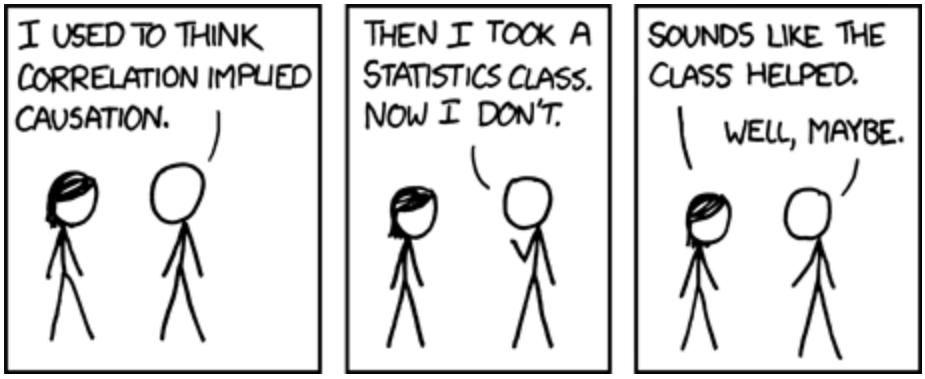

FIGURE 24.1: *"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'" (Image and text copyright by Randall Munroe, used here under a Creative Commons attribution-noncommercial license; see* `http://xkcd.com/552/`. *[[TODO: Excise from the commercial version]])*

expect cotton to burn when we apply flames. We might even say, on the basis of purely statistical evidence, that the world has this habit. But as a matter of pure logic, no amount of evidence about what *did* happen can compel beliefs about what *would* have happened under non-existent circumstances[2]. (For all my *data* shows, all the rags I burn just so happened to be on the verge of spontaneously bursting into flames anyway.) We must supply some counter-factual or causal premise, linking what we see to what we could have seen, to derive causal conclusions.

One of our goals, then, in causal inference will be to make the causal premises as weak and general as possible, thus limiting what we take on faith.

24.2 Causal Graphical Models

We will need a formalism for representing causal relations. It will not surprise you by now to learn that these will be graphical models. We will in fact use DAG models from last time, with "parent" interpreted to mean "directly causes". These will be **causal graphical models**, or **graphical causal models**.[3]

We make the following assumptions.

1. There is some directed acyclic graph G representing the relations of causation among the our variables.

[2]The first person to really recognize this seems to have been the medieval Muslim theologian and anti-philosopher al Ghazali (1100/1997). (See Kogan (1985) for some of the history.) Very similar arguments were made centuries later by Hume (1739); whether there was some line of intellectual descent linking them — that is, any causal connection — I don't know.

[3]Because DAG models have joint distributions which factor according to the graph, we can always write them in the form of a set of equations, as $X_i = f_i(X_{\text{parents}(i)}) + \epsilon_i$, with the catch that the noise ϵ_i is not necessarily independent of X_i's parents. This is what is known, in many of the social sciences, as a **structural equation model**. So those are, strictly, a sub-class of DAG models. They are also often used to represent causal structure.

2. **The Causal Markov condition:** The joint distribution of the variables obeys the Markov property on G.
3. **Faithfulness:** The joint distribution has all of the conditional independence relations implied by the causal Markov property, and *only* those conditional independence relations.

The point of the faithfulness condition is to rule out "conspiracies among the parameters", where, say, two causes of a common effect, which would typically be dependent conditional on that effect, have their impact on the joint effect and their own distributions matched just so exactly that they remain conditionally independent.

24.2.1 Calculating the "effects of causes"

Let's fix two sub-sets of variables in the graph, X_c and X_e. (Assume they don't overlap, and call everything else X_N.) If we want to make a *probabilistic* prediction for X_e's value when X_c takes a particular value, x_c, that's the conditional distribution, $\Pr(X_e \mid X_c = x_c)$, and we saw last time how to calculate that using the graph. Conceptually, this amounts to selecting, out of the whole population or ensemble, the sub-population or sub-ensemble where $X_c = x_c$, and accepting whatever other behavior may go along with that.

Now suppose we want to ask what the effect would be, causally, of setting X_c to a particular value x_c. We represent this by "doing surgery on the graph": we (i) eliminate any arrows coming in to nodes in X_c, (ii) fix their values to x_c, and (iii) calculate the resulting distribution for X_e in the new graph. By steps (i) and (ii), we imagine suspending or switching off the mechanisms which ordinarily set X_c. The other mechanisms in the assemblage are left alone, however, and so step (iii) propagates the fixed values of X_c through them. We are not *selecting* a sub-population, but producing a new one.

If setting X_c to different values, say x_c and x_c', leads to different distributions for X_e, then we say that X_c **has an effect** on X_e — or, slightly redundantly, **has a causal effect** on X_e. Sometimes[4] "the effect of switching from x_c to x_c'" specifically refers to a change in the expected value of X_e, but since profoundly different distributions can have the same mean, this seems needlessly restrictive.[5] If one is interested in average effects of this sort, they are computed by the same procedure.

It is convenient to have a short-hand notation for this procedure of causal conditioning. One more-or-less standard idea, introduced by Judea Pearl, is to introduce a do operator which encloses the conditioning variable and its value. That is,

$$\Pr(X_e \mid X_c = x_c) \tag{24.1}$$

is probabilistic conditioning, or selecting a sub-ensemble from the old mechanisms; but

$$\Pr(X_e \mid do(X_c = x_c)) \tag{24.2}$$

[4] Especially in economics.

[5] Economists are also fond of the horribly misleading usage of talking about "an X effect" or "the effect of X" when they mean the regression coefficient of X. Don't do this.

is causal conditioning, or producing a new ensemble. Sometimes one sees this written as $\Pr\left(X_e \mid X_c \hat{=} x_c\right)$, or even $\Pr\left(X_e \mid \widehat{x_c}\right)$. I am actually fond of the do notation and will use it.

Suppose that $\Pr\left(X_e \mid X_c = x_c\right) = \Pr\left(X_e \mid do(X_c = x_c)\right)$. This would be extremely convenient for causal inference. The conditional distribution on the right is the causal, counter-factual distribution which tells us what would happen if x_c was imposed. The distribution on the left is the ordinary probabilistic distribution we have spent years learning how to estimate from data. When do they coincide?

One situation where they coincide is when X_c contains all the parents of X_e, and none of its descendants. Then, by the Markov property, X_e is independent of all other variables given X_c, and removing the arrows *into* X_c will not change that, or the conditional distribution of X_e given its parents. Doing causal inference for other choices of X_c will demand other conditional independence relations implied by the Markov property. This is the subject of Chapter 25.

24.2.2 Back to Teeth

Let us return to the example of Figure 20.6, and consider the relationship between exposure to asbestos and the staining of teeth. In the model depicted by that figure, the joint distribution factors as

$$\begin{aligned} & p(\text{Yellow teeth}, \text{Smoking}, \text{Asbestos}, \text{Tar in lungs}, \text{Cancer}) \\ = \;& p(\text{Smoking})\, p(\text{Asbestos}) \\ & \times p(\text{Tar in lungs}|\text{Smoking}) \\ & \times p(\text{Yellow teeth}|\text{Smoking}) \\ & \times p(\text{Cancer}|\text{Asbestos}, \text{Tar in lungs}) \end{aligned} \tag{24.3}$$

As we saw, whether or not someone's teeth are yellow (in this model) is unconditionally independent of asbestos exposure, but conditionally *dependent* on asbestos, given whether or not they have cancer. A logistic regression of tooth color on asbestos would show a non-zero coefficient, after "controlling for" cancer. This coefficient would become significant with enough data. The usual interpretation of this coefficient would be to say that the log-odds of yellow teeth increase by so much for each one unit increase in exposure to asbestos, "other variables being held equal".[6] But to see the actual causal effect of increasing exposure to asbestos by one unit, we'd want to compare $p(\text{Yellow teeth}|do(\text{Asbestos} = a))$ to $p(\text{Yellow teeth}|do(\text{Asbestos} = a+1))$, and it's easy to check (Exercise 1) that these two distributions have to be the same. In this case, because asbestos is exogenous, one will in fact get the same result for $p(\text{Yellow teeth}|do(\text{Asbestos} = a)$ and for $p(\text{Yellow teeth}|\text{Asbestos} = a)$.

For a more substantial example, consider Figure 24.2[7] The question of interest here is whether regular brushing and flossing actually prevents heart disease. The

[6] Nothing hinges on this being a logistic regression, similar interpretations are given to all the other standard models.

[7] Based on de Oliveira *et al.* (2010), and the discussion of this paper by Chris Blattman (`http://chrisblattman.com/2010/06/01/does-brushing-your-teeth-lower-cardiovascular-disease/`).

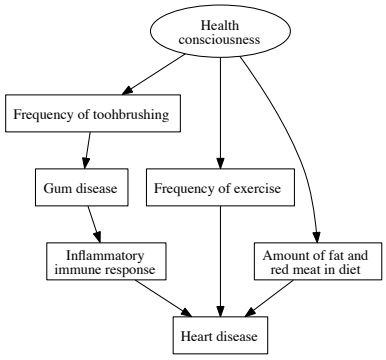

FIGURE 24.2: *Graphical model illustrating hypothetical pathways linking brushing your teeth to not getting heart disease.*

mechanism by which it might do so is as follows: brushing is known to make it less likely for people to get gum disease. Gum disease, in turn, means the gums suffer from constant, low-level inflammation. Persistent inflammation (which can be measured through various messenger chemicals of the immune system) is thought to increase the risk of heart disease. Against this, people who are generally health-conscious are likely to brush regularly, and to take other actions, like regularly exercising and controlling their diets, which also make them less likely to get heart disease. In this case, if we were to manipulate whether people brush their teeth[8], we would shift the graph from Figure 24.2 to Figure 24.3, and we would have

$$p(\text{Heart disease}|\text{Brushing} = b) \neq p(\text{Heart disease}|do(\text{Brushing} = b)) \quad (24.4)$$

[8]Hopefully, by ensuring that everyone brushes, rather than keeping people from brushing.

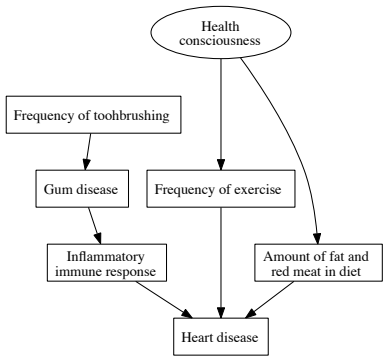

FIGURE 24.3: *The previous graphical model, "surgically" altered to reflect a manipulation (do) of brushing.*

24.3 Conditional Independence and *d*-Separation Revisited

We saw in §20.3 that all distributions which conform to a common DAG share a common set of conditional independence relations. Faithful distributions have *no other* conditional independence relations. These are vital facts for causal inference.

The reason is that while *causal influence* flows one way through the graph, along the directions of arrows from parents to children, *statistical information* can flow in either direction. We can certainly make inferences about an effect from its causes, but we can equally make inferences about causes from their effects. It might be harder to actually do the calculations[9], and we might be left with more uncertainty, but we could do it. As we saw in §20.3, when conditioning on a set of variables S blocks all channels of information flow between X and Y, $X \perp\!\!\!\perp Y | S$. The faithful distributions are the ones where this implication is reversed, where $X \perp\!\!\!\perp Y | S$ implies that S blocks all paths between X and Y. In faithful graphical models, blocking information flow is exactly the same as conditional independence.

This turns out to be the single most important fact enabling causal inference. If we want to estimate the effects of causes, within a given DAG, we need to block off all non-causal channels of information flow. If we want to check whether a given DAG is correct for the variables we have, we need to be able to compare the conditional independence relations implied by the DAG to those supported by the data. If we want to discover the possible causal structures, we have to see which ones imply the conditional independencies supported by the data.

24.4 Further Reading

The two foundational books on graphical causal models are Spirtes *et al.* (2001) and Pearl (2009b). Both are excellent and recommended in the strongest possible terms; but if you had to read just one, I would recommend Spirtes *et al.* (2001). If on the other hand you do not feel up to reading a book at all, then Pearl (2009a) is much shorter, and covers most of the high points. (Also, it's free online.) The textbook by Morgan and Winship (2007, 2015) is much less demanding mathematically, which also means it is less complete conceptually, but it does explain the crucial ideas clearly, simply, and with abundant examples.[10] Lauritzen (1996) has a mathematically rigorous treatment of d-separation (among many other things), but de-emphasizes causality.

[9] Janzing (2007) [[TODO: update refs]] makes the very interesting suggestion that the direction of causality can be discovered by using this — roughly speaking, that if $X|Y$ is much harder to compute than is $Y|X$, we should presume that $X \rightarrow Y$ rather than the other way around.

[10] That textbook also discusses an alternative formalism for counterfactuals, due mainly to Donald B. Rubin. While Rubin has done very distinguished work in causal inference, his formalism is vastly harder to manipulate than are graphical models, but has no more expressive power. (Pearl (2009a) has a convincing discussion of this point, and Richardson and Robins (2013) provides a comprehensive proof that the everything expressible in the counterfactuals formalism can also be expressed with graphical models.) I have accordingly skipped the Rubin formalism here, but good accounts are available in Morgan and Winship (2007, ch. 2), in Rubin's collected papers (Rubin, 2006), and in Imbens and Rubin (2015).

Many software packages for linear structural equation models and path analysis offer options to search for models; these are not, in general, reliable (Spirtes *et al.*, 2001).

Raginsky (2011) provides a fascinating information-theoretic account of graphical causal models and $do()$, in terms of the notion of directed (rather than mutual) information.

[[TODO: historical notes]]

24.5 Exercises

1. Show, for the graphical model in Figure 20.6, that $p(\text{Yellow teeth}|do(\text{Asbestos} = a))$ is always the same as $p(\text{Yellow teeth}|do(\text{Asbestos} = a + 1))$.