14:17 Thursday 9th April, 2015

Chapter 23

Graphical Causal Models

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

23.1 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 23.1). 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 $\mathbf{E}[Y|X=x]$ or $\Pr(Y|X=x)$ 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.

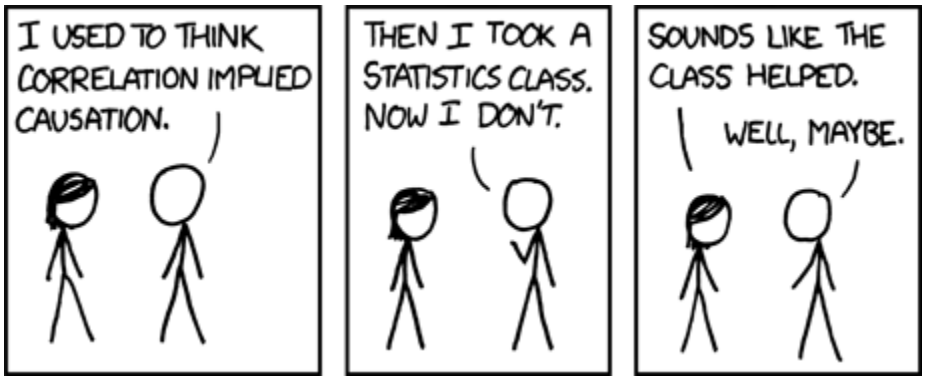

FIGURE 23.1: *"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'" (Image and text copyright by Randall Munroe, used here under a Creative Commons attribution-noncommercial license; see* `http://xkcd.com/552/`. *[[TODO: 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.

23.2 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.

23.2.1 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, $\Pr\left(X_E|X_c = x_c\right)$, 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,

$$\Pr\left(X_E|X_c = x_c\right) \tag{23.1}$$

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

$$\Pr\left(X_E|do(X_c = x_c)\right) \tag{23.2}$$

[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 $\Pr\left(X_E|X_c \hat{=} x_c\right)$, or even $\Pr\left(X_E|\widehat{x_c}\right)$. I am actually fond of the do notation and will use it.

Suppose that $\Pr\left(X_E|X_c = x_c\right) = \Pr\left(X_E|do(X_c = x_c)\right)$. 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 24.

23.2.2 Back to Teeth

Let us return to the example of Figure 22.4, 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

$$\begin{aligned} & p(\text{Yellow teeth}, \text{Smoking}, \text{Asbestos}, \text{Tar in lungs}, \text{Cancer}) \\ = \;& p(\text{Smoking})p(\text{Asbestos}) \\ & \times p(\text{Tar in lungs}|\text{Smoking}) \\ & \times p(\text{Yellow teeth}|\text{Smoking}) \\ & \times p(\text{Cancer}|\text{Asbestos}, \text{Tar in lungs}) \end{aligned} \tag{23.3}$$

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 23.2[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/`).

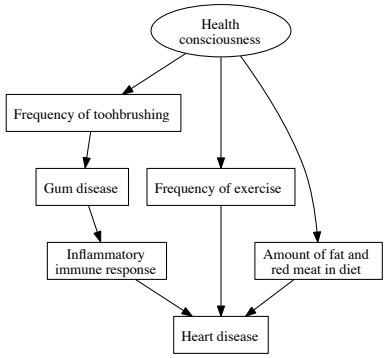

FIGURE 23.2: *Graphical model illustrating hypothetical pathways linking brushing your teeth to not getting heart disease.*

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 23.2 to Figure 23.3, and we would have

$$p(\text{Heart disease}|\text{Brushing} = b) \neq p(\text{Heart disease}|do(\text{Brushing} = b)) \quad (23.4)$$

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

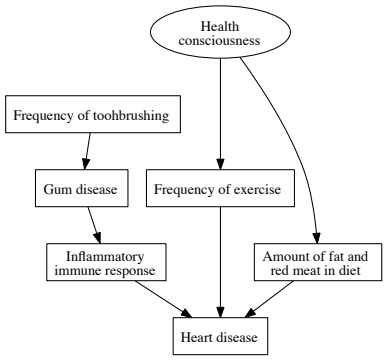

FIGURE 23.3: *The previous graphical model, "surgically" altered to reflect a manipulation (d o) of brushing.*

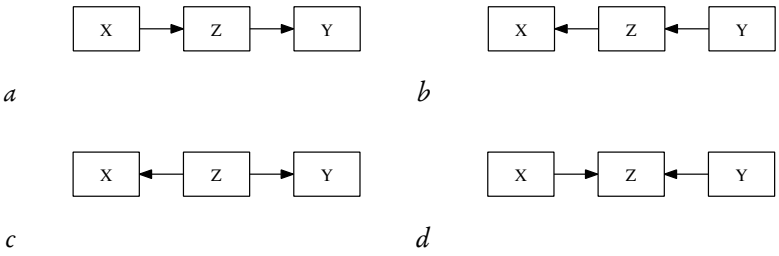

FIGURE 23.4: *Four DAGs for three linked variables. The first two (a and b) are called* **chains***; c is a* **fork***; d is a* **collider***. If these were the whole of the graph, we would have* $X \not\perp\!\!\!\perp Y$ *and* $X \perp\!\!\!\perp Y|Z$*. For the collider, however, we would have* $X \perp\!\!\!\perp Y$ *while* $X \not\perp\!\!\!\perp Y|Z$*.*

23.3 Conditional Independence and *d*-Separation

[[ATTN: Would this section be better in ch. 22?]]

It is clearly very important to us to be able to deduce when two sets of variables are conditionally independent of each other given a third. One of the great uses of DAGs is that they give us a fairly simple criterion for this, in terms of the graph itself. *All* distributions which conform to a given DAG share a common set of conditional independence relations, implied by the Markov property, no matter what their parameters or the form of the distributions. Faithful distributions have *no other* conditional independence relations. Let us think this through.

Our starting point 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.

While we can do inference in either direction across any one edge, we do have to worry about whether we can propagate this information further. Consider the four graphs in Figure 23.4. In every case, we condition on X, which acts as the source of information. In the first three cases, we can (in general) propagate the information from X to Z to Y — the Markov property tells us that Y is independent of its non-descendants given its parents, but in none of those cases does that make X and Y independent. In the last graph, however, what's called a **collider**[10], we cannot propagate the information, because Y has no parents, and X is not its descendant, hence they are independent. We learn about Z from X, but this doesn't tell us anything about Z's other cause, Y.

All of this flips around when we condition on the intermediate variable (Z in Figure 23.4). The chains (Figures 23.4*a* and *b*), conditioning on the intermediate

[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 $X \to Y$ rather than the other way around.

[10] Because two incoming arrows "collide" there.

variable blocks the flow of information from X to Y — we learn nothing more about Y from X and Z than from Z alone, at least not along this path. This is also true of the **fork** (Figure 23.4*c*) — conditional on their common cause, the two effects are uninformative about each other. But in a collider, conditioning on the common effect Z makes X and Y dependent on each other, as we've seen before. In fact, if we don't condition on Z, but do condition on a descendant of Z, we also create dependence between Z's parents.

We are now in a position to work out conditional independence relations. We pick our two favorite variables, X and Y, and condition them both on some third set of variables S. If S **blocks** every undirected path[11] from X to Y, then they must be conditionally independent given S. An unblocked path is also called **active**. A path is active when every variable along the path is active; if even one variable is blocked by S, the whole path is blocked. A variable Z along a path is active, conditioning on S, if

1. Z is a collider along the path, and in S; or,
2. Z is a descendant of a collider, and in S; or
3. Z is not a collider, and not in S.

Turned around, Z is blocked or de-activated by conditioning on S if

1. Z is a non-collider and in S; or
2. Z is collider, and neither Z nor any of its descendants is in S

In words, S blocks a path when it blocks the flow of information by conditioning on the middle node in a chain or fork, and doesn't create dependence by conditioning on the middle node in a collider (or the descendant of a collider). Only *one* node in a path must be blocked to block the whole path. When S blocks *all* the paths between X and Y, we say it **d-separates** them[12]. A collection of variables U is d-separated from another collection V by S if every $X \in U$ and $Y \in V$ are d-separated.

In every distribution which obeys the Markov property, d-separation implies conditional independence[13]. If the distribution is also faithful to the graph, then conditional independence also implies d-separation. In a faithful causal graphical model, then, conditional independence is exactly the same as blocking the flow of information across the graph. This turns out to be the single most important fact enabling causal inference; we will see how that works next time.

23.3.1 D-Separation Illustrated

The discussion of d-separation has been rather abstract, and perhaps confusing for that reason. Figure 23.5 shows a DAG which might make this clearer and more concrete.

[11] Whenever I talk about undirected paths, I mean paths without cycles.

[12] The "d" stands for "directed"

[13] We will not prove this, though I hope I have made it plausible. You can find demonstrations in Spirtes *et al.* (2001); Pearl (2000); Lauritzen (1996).

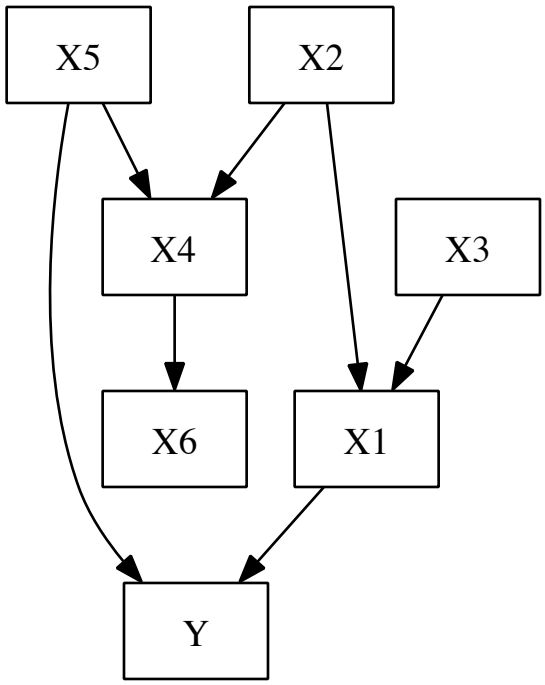

FIGURE 23.5: *Example DAG used to illustrate d-separation.*

If we make the conditioning set S the empty set, that is, we condition on nothing, we "block" paths which pass through colliders. For instance, there are three exogenous variables in the graph, X_2, X_3 and X_5. Because they have no parents, any path from one to another must go over a collider (Exercises 2 and 3). If we do not condition on anything, therefore, we find that the exogenous variables are d-separated and thus independent. Since X_3 is not on any path linking X_2 and X_5, or descended from a node on any such path, if we condition only on X_3, then X_2 and X_5 are still d-separated, so $X_2 \perp\!\!\!\perp X_5 | X_3$. There are two paths linking X_3 to X_5: $X_3 \rightarrow X_1 \leftarrow X_2 \rightarrow X_4 \leftarrow X_5$, and $X_3 \rightarrow X_1 \rightarrow Y \leftarrow X_5$. Conditioning on X_2 (and nothing else) blocks the first path (since X_2 is part of it, but is a fork), and also blocks the second path (since X_2 is not part of it, and Y is a blocked collider). Thus, $X_3 \perp\!\!\!\perp X_5 | X_2$. Similarly, $X_3 \perp\!\!\!\perp X_2 | X_5$ (Exercise 4).

For a somewhat more challenging example, let's look at the relation between X_3 and Y. There are, again, two paths here: $X_3 \rightarrow X_1 \rightarrow Y$, and $X_3 \rightarrow X_1 \leftarrow X_2 \rightarrow X_4 \leftarrow X_5 \rightarrow Y$. If we condition on nothing, the first path, which is a simple chain, is open, so X_3 and Y are d-connected and dependent. If we condition on X_1, we block the first path. X_1 is a collider on the second path, so conditioning on X_1 opens the path there. However, there is a second collider, X_4, along this path, and just conditioning on X_1 does not activate the second collider, so the path as a whole remains blocked.

$$Y \not\perp\!\!\!\perp X_3 \tag{23.5}$$
$$Y \perp\!\!\!\perp X_3 | X_1 \tag{23.6}$$

To activate the second path, we can condition on X_1 and either X_4 (a collider along that path) or on X_6 (a descendant of a collider) or on both:

$$Y \not\perp\!\!\!\perp X_3 | X_1, X_4 \tag{23.7}$$
$$Y \not\perp\!\!\!\perp X_3 | X_1, X_6 \tag{23.8}$$
$$Y \not\perp\!\!\!\perp X_3 | X_1, X_4, X_6 \tag{23.9}$$

Conditioning on X_4 and/or X_6 does not activate the $X_3 \rightarrow X_1 \rightarrow Y$ path, but it's enough for there to be one active path to create dependence.

To block the second path again, after having opened it in one of these ways, we can condition on X_2 (since it is a fork along that path, and conditioning on a fork blocks it), or on X_5 (also a fork), or on both X_2 and X_5. So

$$Y \perp\!\!\!\perp X_3 | X_1, X_2 \tag{23.10}$$
$$Y \perp\!\!\!\perp X_3 | X_1, X_5 \tag{23.11}$$
$$Y \perp\!\!\!\perp X_3 | X_1, X_2, X_5 \tag{23.12}$$
$$Y \perp\!\!\!\perp X_3 | X_1, X_2, X_4 \tag{23.13}$$
$$Y \perp\!\!\!\perp X_3 | X_1, X_2, X_6 \tag{23.14}$$
$$Y \perp\!\!\!\perp X_3 | X_1, X_2, X_5, X_6 \tag{23.15}$$

etc., etc.

Let's look at the relationship between X_4 and Y. X_4 is not an ancestor of Y, or a descendant of it, but they do share common ancestors, X_5 and X_2. Unconditionally,

Y and X_4 are dependent, both through the path going $X_4 \leftarrow X_5 \rightarrow Y$, and through that going $X_4 \leftarrow X_2 \rightarrow X_1 \rightarrow Y$. Along both paths, the exogenous variables are forks, so *not* conditioning on them leaves the path unblocked. X_4 and Y become d-separated when we condition on X_5 and X_2.

X_6 and X_3 have no common ancestors. Unconditionally, they should be independent, and indeed they are: the two paths are $X_6 \leftarrow X_4 \leftarrow X_2 \rightarrow X_1 \leftarrow X_3$, and $X_6 \leftarrow X_4 \leftarrow X_5 \rightarrow Y \leftarrow X_1 \leftarrow X_3$. Both paths contain a single collider (X_1 and Y, respectively), so if we do not condition on them the paths are blocked and X_6 and X_3 are independent. If we condition on either Y or X_1 (or both), however, we unblock the paths, and X_6 and X_3 become d-connected, hence dependent. To get back to d-separation while conditioning on Y, we must also condition on X_4 or X_5, or both. To get d-separation while conditioning on X_1, we must also condition on X_4, or on X_2, or on X_4 and X_2. If we condition on both X_1 and Y and want d-separation, we could just add conditioning on X_4, or we could condition on X_2 and X_5, or all three.

If the abstract variables are insufficiently concrete, consider reading them as follows:

$$
\begin{aligned}
Y &\Leftrightarrow \text{Grade in this class}\\
X_1 &\Leftrightarrow \text{Effort spent on this class}\\
X_2 &\Leftrightarrow \text{Enjoyment of statistics}\\
X_3 &\Leftrightarrow \text{Workload this term}\\
X_4 &\Leftrightarrow \text{Quality of work in linear regression class}\\
X_5 &\Leftrightarrow \text{Amount learned in linear regression class}\\
X_6 &\Leftrightarrow \text{Grade in linear regression}
\end{aligned}
$$

Pretending, for the sake of illustration, that this is accurate, how heavy your workload is this semester (X_3) would predict, or rather retrodict, your grade in linear regression last semester (X_6), once we control for how much effort you put into this class (X_1). Changing your workload this semester would not, however, reach backwards in time to raise or lower your grade in regression.

23.3.2 Linear Graphical Models and Path Coefficients

We began our discussion of graphical models with factor analysis as our starting point. Factor models are a special case of linear (directed) graphical models, a.k.a. path models[14] As with factor models, in the larger class we typically center all the variables (so they have expectation zero) and scale them (so they have variance 1). In factor models, the variables were split into two sets, the factors and the observables, and all the arrows went from factors to observables. In the more general case, we do not necessarily have this distinction, but we still assume the arrows from a directed acyclic graph. The conditional expectation of each variable is a linear combination of the values of its parents:

$$
\mathbf{E}\left[X_i | X_{\text{parents}(i)}\right] = \sum_{j \in \text{parents}(i)} w_{ji} X_j \tag{23.16}
$$

[14]Some people use the phrase "structural equation models" for such models exclusively.

just as in a factor model. In a factor model, the coefficients w_{ji} were the factor loadings. More generally, they are called **path coefficients**.

The path coefficients determine all of the correlations between variables in the model. To find the correlation between X_i and X_j, we proceed as follows:

- Find all of the undirected paths between X_i and X_j.
- Discard all of the paths which go through colliders.
- For each remaining path, multiply all the path coefficients along the path.
- Sum up these products over paths.

These rules were introduced by the great geneticist and mathematical biologist Sewall Wright in the early 20th century, in a series of papers culminating in Wright (1934)[15] These "Wright path rules" often seem mysterious, particularly the bit where paths with colliders are thrown out. But from our perspective, we can see that what Wright is doing is finding all of the *unblocked* paths between X_i and X_j. Each path is a channel along which information (here, correlation) can flow, and so we add across channels.

It is frequent, and customary, to assume that all of the variables are Gaussian. (We saw this in factor models as well.) With this extra assumption, the joint distribution of all the variables is a multivariate Gaussian, and the correlation matrix (which we find from the path coefficients) gives us the joint distribution.

If we want to find conditional correlations, $\mathrm{corr}(X_i, X_j | X_k, X_l, \ldots)$, we still sum up over the unblocked paths. If we have avoided conditioning on colliders, then this is just a matter of dropping the now-blocked paths from the sum. If on the other hand we have conditioned on a collider, that path *does* become active (unless blocked elsewhere), and we in fact need to modify the path weights. Specifically, we need to work out the correlation induced between the two parents of the collider, by conditioning on that collider. This can be calculated from the path weights, and some fairly tedious algebra[16]. The important thing is to remember that the rule of d-separation still applies, and that conditioning on a collider can create correlations.

23.3.3 Positive and Negative Associations

We say that variables X and Y are **positively associated** if increasing X predicts, on average, an increase in Y, and vice versa[17]; if increasing X predicts a decrease in Y, then they are **negatively associated**. If this holds when conditioning out other variables, we talk about positive and negative partial associations. Heuristically, positive association means positive correlation in the neighborhood of any given x, though the magnitude of the positive correlation need not be constant. Note that not all dependent variables have to have a definite sign for their association.

[15] That paper is now freely available online, and worth reading. See also `http://www.ssc.wisc.edu/soc/class/soc952/Wright/wright_biblio.htm` for references to, and in some cases copies of, related papers by Wright.

[16] See for instance Li *et al.* (1975).

[17] I.e., if $\frac{d\mathbf{E}[Y|X=x]}{dx} \geq 0$

We can multiply together the signs of positive and negative partial associations along a path in a graphical model, the same we can multiply together path coefficients in a linear graphical model. Paths which contain (inactive!) colliders should be neglected. If all the paths connecting X and Y have the same sign, then we know that over-all association between X and Y must have that sign. If different paths have different signs, however, then signs alone are not enough to tell us about the over-all association.

If we are interested in conditional associations, we have to consider whether our conditioning variables block paths or not. Paths which are blocked by conditioning should be dropped from consideration. If a path contains an activated collider, we need to include it, but we reverse the sign of one arrow into the collider. That is, if $X \stackrel{+}{\rightarrow} Z \stackrel{+}{\leftarrow} Y$, and we condition on Z, we need to replace one of the plus signs with a $-$ sign, because the two parents now have an over-all negative association.[18] If on the other hand one of the incoming arrows had a positive association and the other was negative, we need to flip one of them so they are both positive or both negative; it doesn't matter which, since it creates a positive association between the parents[19].

[[TODO: Write out formal proofs as appendix]]

23.4 Independence, Conditional Independence, and Information Theory

Take two random variables, X and Y. They have some joint distribution, which we can write $p(x,y)$. (If they are both discrete, this is the joint probability mass function; if they are both continuous, this is the joint probability density function; if one is discrete and the other is continuous, there's still a distribution, but it needs more advanced tools.) X and Y each have marginal distributions as well, $p(x)$ and $p(y)$. $X \perp\!\!\!\perp Y$ if and only if the joint distribution is the product of the marginals:

$$X \perp\!\!\!\perp Y \Leftrightarrow p(x,y) = p(x)p(y) \tag{23.17}$$

We can use this observation to measure how dependent X and Y are. Let's start with the log-likelihood ratio between the joint distribution and the product of marginals:

$$\log \frac{p(x,y)}{p(x)p(y)} \tag{23.18}$$

This will always be exactly 0 when $X \perp\!\!\!\perp Y$. We use its average value as our measure of dependence:

$$I[X;Y] \equiv \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{23.19}$$

[18] If both smoking and asbestos are positively associated with lung cancer, and we know the patient does not have lung cancer, then high levels of smoking must be compensated for by low levels of asbestos, and vice versa.

[19] If yellow teeth are positively associated with smoking and negatively associated with dental insurance, and we know the patient does not have yellow teeth, then high levels of smoking must be compensated for by excellent dental care, and conversely poor dental care must be compensated for by low levels of smoking.

(If the variables are continuous, replace the sum with an integral.) Clearly, if $X \perp\!\!\!\perp Y$, then $I[X;Y] = 0$. One can show[20] that $I[X;Y] \geq 0$, and that $I[X;Y] = 0$ implies $X \perp\!\!\!\perp Y$. The quantity $I[X;Y]$ is clearly symmetric between X and Y. Less obviously, $I[X;Y] = I[f(X);g(Y)]$ whenever f and g are invertible functions. This **coordinate-freedom** means that $I[X;Y]$ measures *all forms* of dependence, not just linear relationships, like the ordinary (Pearson) correlation coefficient, or monotone dependence, like the rank (Spearman) correlation coefficient. In information theory, $I[X;Y]$ is called the **mutual information**, or **Shannon information**, between X and Y. So we have the very natural statement that random variables are independent just when they have no information about each other.

There are (at least) two ways of giving an operational meaning to $I[X;Y]$. One, the original use of the notion, has to do with using knowledge of Y to improve the efficiency with which X can be encoded into bits (Shannon, 1948; Cover and Thomas, 2006). While this is very important — it's literally transformed the world since 1945 — it's not very statistical. For statisticians, what matters is that if we test the hypothesis that X and Y are independent, with joint distribution $p(x)p(y)$, against the hypothesis that they dependent, with joint distribution $p(x,y)$, then the mutual information controls the error probabilities of the test. To be exact, if we fix any power we like (90%, 95%, 99.9%, ...), the size or type I error rate α_n, of the best possible test shrinks exponentially with the number of IID samples n, and the rate of exponential decay is precisely $I[X;Y]$ (Kullback, 1968, §4.3, theorem 4.3.2):

$$\lim_{n\to\infty} -\frac{1}{n}\log \alpha_n \leq I[X;Y] \tag{23.20}$$

So positive mutual information means dependence, and the magnitude of mutual information tells us about how detectable the dependence is[21].

Suppose we conditioned X and Y on a third variable (or variables) Z. For each realization z, we can calculate the mutual information,

$$I[X;Y|Z=z] \equiv \sum_{x,y} p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{23.21}$$

And we can average over z,

$$I[X;Y|Z] \equiv \sum_{z} p(z) I[X;Y|Z=z] \tag{23.22}$$

This is the **conditional mutual information**. It will not surprise you at this point to learn that $X \perp\!\!\!\perp Y|Z$ if and only if $I[X;Y|Z] = 0$. The magnitude of the conditional mutual information tells us how easy it is to detect conditional dependence.

[20] Using the same type of convexity argument ("Jensen's inequality") we used §21.2.1 for understanding why the EM algorithm works.

[21] Symmetrically, if we follow the somewhat more usual procedure of fixing a type I error rate α, the type II error rate β_n (= 1−power) also goes to zero exponentially, and the exponential rate is $\sum_{x,y} p(x)p(y)\log\frac{p(x)p(y)}{p(x,y)}$, a quantity called the "lautam information" (Palomar and Verdú, 2008). (For proofs of the exponential rate, see Palomar and Verdú (2008, p. 965), following Kullback (1968, §4.3, theorem 4.3.3).)

23.5 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.[22] Lauritzen (1996) has a mathematically rigorous treatment of d-separation (among many other things), but de-emphasizes causality.

Linear path models have a very large literature, going back to the early 20th century; see references in the previous chapter. 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).

On information theory (§23.4), the best book is Cover and Thomas (2006) by a large margin. 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]]

23.6 Exercises

1. Show, for the graphical model in Figure 22.4, that $p(\text{Yellow teeth}|do(\text{Asbestos} = a))$ is always the same as $p(\text{Yellow teeth}|do(\text{Asbestos} = a+1))$.

2. Find all the paths between the exogenous variables in Figure 23.5, and verify that every such path goes through at least one collider .

3. Is it true that in any DAG, every path between exogenous variables must go through at least one collider, or descendant of a collider? Either prove it or construct a counter-example in which it is not true. Does the answer change we say "go through at least one collider", rather than "collider or descendant of a collider"? .

4. Prove that $X_2 \perp\!\!\!\perp X_3 | X_5$ in Figure 23.5.

[22] This textbook also discusses an alternative formalism for counterfactuals, due to Donald 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), and in Rubin's collected papers (Rubin, 2006).