Causal Inference

36-350, Data Mining, Fall 2009

4 December 2009

Contents

There are two problems which are both known as "causal inference":

1. Given the causal structure of a system, estimate the effects the variables have on each other.
2. Given data about a system, find its causal structure.

The first problem is easier, so we'll begin with it.

1 Estimating Causal Effects with Known Structure

Suppose we want to estimate the causal effect of X on Y, $\Pr(Y|do(X = x))$. If we can actually manipulate the system, then the *statistical* problem is trivial: set X to x, measure Y, and repeat often enough to get an estimate of the distribution. (As my mother says, "Why *think* when you can just do the experiment?") So suppose we can't do the experiment. We can however get the joint distribution $\Pr(X, Y, Z)$ for some collection of **covariates** Z, and we know the causal graph. Is this enough to determine $\Pr(Y|do(X = x))$? That is, does the joint distribution **identify** the causal effect?

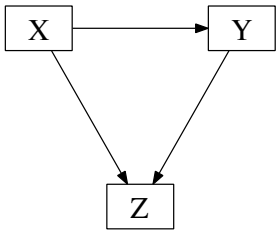

Figure 1: "Controlling for" additional variables can introduce bias into estimates of causal effects. Here the effect of X on Y is directly identifiable, $\Pr(Y|do(X=x)) = \Pr(Y|X=x)$. If we also condition on Z however, because it is a common *effect* of X and Y, we'd get $\Pr(Y|X=x, Z=z) \neq \Pr(Y|X=x)$. In fact, even if there were no arrow from X to Y, conditioning on Z would make Y depend on X.

The answer is "yes" when the covariates Z contain all the other relevant variables. The inferential problem is then trivial again, or at least no worse than any other statistical estimation problem. In fact, if we know the causal graph and get to observe all the variables, then we could (in principle) just use our favorite non-parametric conditional density estimate at each node in the graph, with its parent variables as the inputs and its own variable as the response. Multiplying conditional distributions together gives the whole distribution of the graph, and we can get any causal effects we want by surgery. If we're willing to assume a bit more, we can get away with just using non-parametric regression or even just an additive model at each node. Assuming yet more, we could use parametric models at each node; the linear-Gaussian assumption is (alas) very popular.

If some variables are *not* observed, then the issue of which causal effects are observationally identifiable is considerably trickier. Apparently subtle changes in which variables are available to us and used can have profound consequences.

The basic principle underlying all considerations is that we would like to condition on adequate **control** variables, which will block paths linking X and Y *other than* those which would exist in the surgically-altered graph where all paths into X have been removed. If other unblocked paths exist, then there is some confounding of the causal effect of X on Y with their mutual association with third parties. Just conditioning on everything possible does *not* give us adequate control, or even necessarily bring us closer to it (Figure 1 and Exercise 1).

There are two main sufficient criteria we can use to get adequate control;

they are called the **back-door criterion** and the **front-door criterion**.

If we want to know the effect of X on Y and have a set of variables S as the control, then S satisfies the back-door criterion if (i) S blocks every path from X to Y that has an arrow *into* X ("blocks the back door"), and (ii) no node in S is a descendant of X. Then

$$\Pr(Y|do(X=x)) = \sum_s \Pr(Y|X=x, S=s)\Pr(S=s) \tag{1}$$

Notice that all the items on the right-hand side are observational conditional probabilities, not counterfactuals.

On the other hand, S satisfies the front-door criterion when (i) S blocks all directed paths from X to Y, (ii) there are no unblocked back-door paths from X to S, and (iii) X blocks all back-door paths from S to Y. Then

$$\Pr(Y|do(X=x)) = \sum_s \Pr(S=s|X=x) \sum_{x'} \Pr(Y|X=x', S=s)\Pr(X=x') \tag{2}$$

A natural reaction to the front-door criterion is "Say what?", but it becomes more comprehensible if we take apart its. Because, by clause (i), S blocks all *directed* paths from X to Y, any *causal* dependence of Y on X must be mediated by a dependence of Y on S:

$$\Pr(Y|do(X=x)) = \sum_s \Pr(Y|do(S=s))\Pr(S=s|do(X=x))$$

Clause (ii) says that we can estimate the effect of X on S directly,

$$\Pr(S=s|do(X=x)) = \Pr(S=s|X=x) \ .$$

Clause (iii) say that X satisfies the back-door criterion for estimating the effect of S on Y, and the inner sum in Eq. 2 is just the back-door estimate (Eq. 1) of $\Pr(Y|do(S=s))$. So really we *are* using the back door criterion. (See Figure 2.)

Both the back-door and front-door criteria are *sufficient* for estimating causal effects from probabilistic distributions, but are not *necessary*. Necessary and sufficient conditions for the identifiablity of causal effects are in principle possible but don't have a nice snappy form (Pearl, 2009, §§3.4–3.5). A necessary condition for *un*-identifiability, however, is the presence of an unblockable back-door path from X to Y. However, this is not sufficient for lack of identification — we might, for instance, be able to use the front door criterion, as in Figure 2.

When identification — that is, adequate control of confounding — is not possible, it may still be possible to *bound* causal effects. That is, even if we can't say exactly that $\Pr(Y|do(X=x))$ must be, we can still say it has to fall within a certain (non-trivial!) range of possibilities. The development of bounds for non-identifiable quantities, what's sometimes called **partial identification**, is an active area of research, which I think is very likely to work its way back into data-mining; the best introduction is probably Manski (2007).

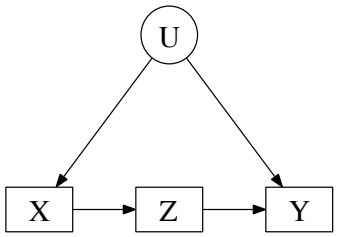

Figure 2: Illustration of the front-door criterion, after Pearl (2009, Figure 3.5). X, Y and Z are all observed, but U is an unobserved common cause of both X and Y. $X \leftarrow U \rightarrow Y$ is a back-door path confounding the effect of X on Y with their common cause. However, all of the effect of X on Y is mediated through X's effect on Z. Z's effect on Y is, in turn, confounded by the back-door path $Z \leftarrow X \leftarrow U \rightarrow Y$, but X blocks this path. So we can use back-door adjustment to find $\Pr(Y|do(Z=z))$, and directly find $\Pr(Z|do(X=x)) = \Pr(Z|X=x)$, and putting these together gives $\Pr(Y|do(X=x))$.

2 Discovering Causal Structure

2.1 Causal Discovery with Known Variables

Causal discovery is silly with just one variable, and too hard with just two for us.[1]

So let's start with three variables, X, Y and Z. By testing for independence and conditional independence, we could learn that there had to be edges between X and Y and Y and Z, but not between X and Z.[2] But conditional independence is a symmetric relationship, so how could we **orient** those edges, give them direction? Well, there are only four possible directed graphs corresponding to that undirected graph:

- $X \rightarrow Y \rightarrow Z$ (a chain);
- $X \leftarrow Y \leftarrow Z$ (the other chain);

[1] But see Janzing (2007); Hoyer *et al.* (2009) for some ideas on how you could do it if you're willing to make some extra assumptions. The basic idea of these papers is that the distribution of effects given causes should be simpler, in some sense, than the distribution of causes given effects.

[2] Remember that an edge between X and Y means that either X is a parent of Y, $X \rightarrow Y$, or Y is a parent of X, $X \leftarrow Y$. Either way, the two variables will be dependent no matter what collection of other variables we might condition on. If $X \perp\!\!\!\perp Y|S$ for some set of variables S, then, and only then, is there no edge between X and Y.

- $X \leftarrow Y \rightarrow Z$ (a fork on Y);
- $X \rightarrow Y \leftarrow Z$ (a collision at Y)

With the fork or either chain, we have $X \perp\!\!\!\perp Z|Y$. On the other hand, with the collider we have $X \not\perp\!\!\!\perp Z|Y$. (This is where the assumption of faithfulness comes in.) Thus $X \not\perp\!\!\!\perp Z|Y$ if and only if there is a collision at Y. By testing for this conditional independence, we can either definitely orient the edges, or rule out an orientations. If $X - Y - Z$ is just a subgraph of a larger graph, we can still identify it as a collider if $X \not\perp\!\!\!\perp Z|\{Y, S\}$ for *all* collections of nodes S (not including X and Z themselves, of course).

With more nodes and edges, we can **induce** more orientations of edges by consistency with orientations we get by identifying colliders. For example, suppose we know that X, Y, Z is either a chain or a fork on Y. If we learn that $X \rightarrow Y$, then the triple *cannot* be a fork, and must be the chain $X \rightarrow Y \rightarrow Z$. So orienting the $X - Y$ edge induces an orientation of the $Y - Z$ edge. We can also sometimes orient edges through background knowledge; for instance we might know that Y comes later in time than X, so if there is an edge between them it *cannot* run from Y to X.[3] We can eliminate other edges based on similar sorts of background knowledge: men tend to be heavier than women, but changing weight does not change sex, so there can't be an edge (or even a directed path!) from weight to sex.

Orienting edges is the core of the basic causal discovery procedure, the SGS algorithm (Spirtes *et al.*, 2001, §5.4.1, p. 82). This assumes:

1. The data-generating distribution has the causal Markov property on a graph G.
2. The data-generating distribution is faithful to G.
3. Every member of the population has the same distribution.
4. All relevant variables are in G.
5. There is only *one* graph G to which the distribution is faithful.

Abstractly, the algorithm works as follows:

- Start with a complete undirected graph on all variables.

[3]Some have argued, or at least entertained the idea, that the logic here is backwards: rather than order in time constraining causal relations, causal order *defines* time order. (Versions of this idea are discussed by, inter alia, Russell (1927); Wiener (1961); Reichenbach (1956); Pearl (2009); Janzing (2007) makes a related suggestion). Arguably then using order in time to orient edges in a causal graph begs the question, or commits the fallacy of *petitio principii*. But of course every syllogism does, so this isn't a distinctively *statistical* issue. (Take the classic: "All men are mortal; Socrates is a man; therefore Socrates is mortal." How can we know that *all* men are mortal until we know about the mortality of this particular man, Socrates? Isn't this just like asserting that tomatoes and peppers must be poisonous, because they belong to the nightshade family of plants, all of which are poisonous?) While these philosophical issues are genuinely fascinating, this footnote has gone on long enough, and it is time to return to the main text.

- For each pair of variables, see if conditioning on some set of variables makes them conditionally independent; if so, remove their edge.
- Identify all colliders by checking for conditional dependence; orient the edges of colliders.
- Try to orient undirected edges by consistency with already-oriented edges; do this recursively until no more edges can be oriented.

Pseudo-code is in the appendix.

Call the result of the SGS algorithm $\widehat{G}$. If all of the assumptions above hold, and the algorithm is correct in its guesses about when variables are conditionally independent, then $\widehat{G} = G$. In practice, of course, conditional independence guesses are really statistical tests based on finite data, so we should write the output as $\widehat{G}_n$, to indicate that it is based on only n samples. If the conditional independence test is consistent, then

$$\lim_{n\to\infty} \Pr\left(\widehat{G}_n \neq G\right) = 0$$

In other words, the SGS algorithm converges in probability on the correct causal structure; it is consistent for all graphs G. Of course, at finite n, the probability of error — of having the wrong structure — is (generally!) not zero, but this just means that, like any statistical procedure, we cannot be absolutely certain that it's not making a mistake.

One consequence of the independence tests making errors on finite data can be that we fail to orient some edges — perhaps we missed some colliders. These unoriented edges in $\widehat{G}_n$ can be thought of as something like a confidence region — they have *some* orientation, but multiple orientations are all compatible with the data.[4] As more and more edges get oriented, the confidence region shrinks.

If the fifth assumption above fails to hold, then there are multiple graphs G to which the distribution is faithful. This is just a more complicated version of the difficulty of distinguishing between the graphs $X \to Y$ and $X \leftarrow Y$. All the graphs in this **equivalence class** may have some arrows in common; in that case the SGS algorithm will identify those arrows. If some edges differ in orientation across the equivalence class, SGS will not orient them, even in the limit. In terms of the previous paragraph, the confidence region never shrinks to a single point, just because the data doesn't provide the information needed to do this.

If there *are* unmeasured relevant variables, we can get not just unoriented edges, but actually arrows pointing in both directions. This is an excellent sign that some basic assumption is being violated.

The SGS algorithm is statistically consistent, but very computationally inefficient; the number of tests it does grows exponentially in the number of variables p. This is the worst-case complexity for *any* consistent causal-discovery procedure, but this algorithm just proceeds immediately to the worst case, not

[4] I say "multiple orientations" rather than "all orientations", because picking a direction for one edge might induce an orientation for others.

taking advantage of any possible short-cuts. A refinement, called the PC algorithm, tries to minimize the number of conditional independence tests performed (Spirtes *et al.*, 2001, §5.4.2, pp. 84–88). There is actually an implementation of the PC algorithm in R (`PCalg` on CRAN), but it assumes linear-Gaussian models (Kalisch and Bühlmnann, 2007).

2.2 Causal Discovery with Hidden Variables

Suppose that the set of variables we measure is *not* causally sufficient. Could we at least discover this? Could we possibly get hold of *some* of the causal relationships? Algorithms which can do this exist (e.g., the CI and FCI algorithms of Spirtes *et al.* (2001, ch. 6)), but they require considerably more graph-fu. The results of these algorithms can succeed in removing *some* edges between observable variables, and definitely orienting some of the remaining edges. If there are actually no latent common causes, they end up acting like the SGS or PC algorithms.

There is not, so far as I know, any implementation of the CI or FCI algorithms in R. The FCI and PC algorithms, along with some other, related procedures, are implemented in the stand-alone Java program `Tetrad` (`http://www.phil.cmu.edu/projects/tetrad/`). It would be a Good Thing if someone were to re-implement these algorithms in R.

2.3 Note on Conditional Independence Tests

The abstract algorithms for causal discovery assume the existence of consistent tests for conditional independence. The implementations known to me assume either that variables are discrete (so that one can basically use the χ^2 test), or that they are continuous, Gaussian, and linearly related (so that one can test for vanishing partial correlations). It bears emphasizing that these restrictions are *not* essential. As soon as you have a consistent independence test, you are, in principle, in business. In particular, consistent *non-parametric* tests of conditional independence would work perfectly well. An interesting example of this is the paper by Chu and Glymour (2008), on finding causal models for the time series, assuming additive but non-linear models.

2.4 Limitations on Consistency of Causal Discovery

There are some important limitations to causal discovery algorithms (Spirtes *et al.*, 2001, §12.4). They are *universally* consistent: for all causal graphs G,[5]

$$\lim_{n\to\infty} \Pr\left(\widehat{G}_n \neq G\right) = 0 \tag{3}$$

The probability of getting the graph wrong can be made arbitrarily small by using enough data. However, this says nothing about *how much* data we need

[5]If the true distribution is faithful to multiple graphs, then we should read G as their common graph pattern, which has some undirected edges.

to achieve a given level of confidence, i.e., the *rate* of convergence. *Uniform* consistency would mean that we could put a bound on the probability of error as a function of n which did not depend on the true graph G. Robins *et al.* (2003) proved that *no* uniformly-consistent causal discovery algorithm can exist. The issue, basically, is that the Adversary could make the convergence in Eq. 3 arbitrarily slow by selecting a distribution which, while faithful to G, came *very close* to being unfaithful, making some of the dependencies implied by the graph arbitrarily small. For any given dependence strength, there's some amount of data which will let us recognize it with high confidence, but the Adversary can make the required data size as large as he likes by weakening the dependence, without ever setting it to zero.[6]

The upshot is that so *uniform*, *universal* consistency is out of the question; we can be *universally* consistent, but without a uniform rate of convergence; or we can converge *uniformly*, but only on some less-than-universal class of distributions. These might be ones where all the dependencies which do exist are not too weak (and so not too hard to learn reliably from data), or the number of true edges is not too large (so that if we haven't seen edges yet they probably don't exist; Janzing and Herrmann, 2003; Kalisch and Bühlmnann, 2007).

It's worth emphasizing that the Robins *et al.* (2003) no-uniform-consistency result applies to *any* method of discovering causal structure from data. Invoking human judgment, Bayesian priors over causal structures, etc., etc., won't get you out of it.

[6]Roughly speaking, if X and Y are dependent given Z, the probability of missing this conditional dependence with a sample of size n should go to zero like $O(2^{-nI[X;Y|Z]})$, I being mutual information. To make this probability equal to, say, α we thus need $n = O(-\log \alpha / I)$ samples. The Adversary can thus make n extremely large by making I very small, yet positive.

3 Exercises

Not to hand in.

1. Take the model in Figure 1. Suppose that $X \sim \mathcal{N}(0, \sigma_X^2)$, $Y = \alpha X + \epsilon$ and $Z = \beta_1 X + \beta_2 Y + \eta$, where ϵ and η are mean-zero Gaussian noise. Set this up in R and run regress Y twice, once on X alone and once on X and Z. Can you find any values of the parameters where the coefficient of X in the second regression is even approximately equal to α? (It's possible to solve this problem exactly through linear algebra instead.)

2. Take the model in Figure 2 and parameterize it as follows: $U \sim \mathcal{N}(0, 1)$, $X = \alpha_1 U + \epsilon$, $Z = \beta X + \eta$, $Y = \gamma Z + \alpha_2 U + \xi$, where ϵ, η, ξ are independent Gaussian noises. If you regress Y on Z, what coefficient do you get? If you regress Y on Z and X? If you do a back-door adjustment for X? (Approach this either analytically or through simulation, as you like.)

3. Continuing in the set-up of the previous problem, what coefficient do you get for X when you regress Y on Z and X? Now compare this to the front-door adjustment for the effect of X on Y.

A Pseudocode for the SGS Algorithm

When you see a loop, assume that it gets entered at least once. "Replace" in the sub-functions always refers to the input graph.

```
SGS = function(set of variables V) {
    Ĝ = colliders(prune( complete undirected graph on V))
    until (Ĝ == G′) {
          Ĝ = G′
          G′ = orient(Ĝ)
    }
    return(Ĝ)
}

prune = function(G) {
    for each A, B ∈ V {
        for each S ⊆ V \ {A, B} {
            if A⫫B|S { G = G \ (A − B) }
        }
    }
    return(G)
}

collliders = function(G) {
    for each (A − B) ∈ G {
        for each (B − C) ∈ G {
            if (A − C) ∉ G {
               collision = TRUE
               for each S ⊂ B ∩ V \ {A, C} {
                   if A⫫C|S { collision = FALSE }
               }
               if (collision) { replace (A − B) with (A → B), (B − C) with (B ← C) }
            }
        }
    }
    return(G)
}

orient = function(G) {
    if ((A → B) ∈ G & (B − C) ∈ G & (A − C) ∉ G) { replace (B − C) with (B → C) }
    if ((directed path from A to B)∈ G & (A − B) ∈ G) { replace (A − B) with (A → B) }
    return(G)
}
```

References

Chu, Tianjiao and Clark Glymour (2008). "Search for Additive Nonlinear Time Series Causal Models." *Journal of Machine Learning Research*, **9**: 967–991. URL `http://jmlr.csail.mit.edu/papers/v9/chu08a.html`.

Hoyer, Patrik O., Domink Janzing, Joris Mooij, Jonas Peters and Bernhard Schölkopf (2009). "Nonlinear causal discovery with additive noise models." In *Advances in Neural Information Processing Systems 21 [NIPS 2008]* (D. Koller and D. Schuurmans and Y. Bengio and L. Bottou, eds.), pp. 689–696. Cambridge, Massachusetts: MIT Press. URL `http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf`.

Janzing, Dominik (2007). "On causally asymmetric versions of Occam's Razor and their relation to thermodynamics." E-print, arxiv.org. URL `http://arxiv.org/abs/0708.3411`.

Janzing, Dominik and Daniel Herrmann (2003). "Reliable and Efficient Inference of Bayesian Networks from Sparse Data by Statistical Learning Theory." Electronic preprint. URL `http://arxiv.org/abs/cs.LG/0309015`.

Kalisch, Markus and Peter Bühlmnann (2007). "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm." *Journal of Machine Learning Research*, **8**: 616–636. URL `http://jmlr.csail.mit.edu/papers/v8/kalisch07a.html`.

Manski, Charles F. (2007). *Identification for Prediction and Decision*. Cambridge, Massachusetts: Harvard University Press.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press, 2nd edn.

Reichenbach, Hans (1956). *The Direction of Time*. Berkeley: University of California Press. Edited by Maria Reichenbach.

Robins, James M., Richard Scheines, Peter Spirtes and Larry Wasserman (2003). "Uniform Consistency in Causal Inference." *Biometrika*, **90**: 491–515. URL `http://www.stat.cmu.edu/tr/tr725/tr725.html`.

Russell, Bertrand (1927). *The Analysis of Matter*. International Library of Philosophy, Psychology and Scientific Method. London: K. Paul Trench, Trubner and Co. Reprinted New York: Dover Books, 1954.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.

Wiener, Norbert (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.