Lecture 24, Causal Discovery

36-402, Advanced Data Analysis

21 April 2011

Contents

Last time, we looked at the problem of estimating causal effects within a known graphical causal model — essentially the problem of removing confounding. Today, at last, we get at the problem of how to find the right graph in the first place. As always, we presume that there *is* some directed acyclic graph which adequately represents the systematic interactions among the variables.

First, as a warm-up, we look at testing the implications of different DAG models, and so comparing them.

1 Testing DAGs

As seen in the homework, if we have multiple contending DAGs, we would like to focus our inference on telling which one is right (if any of them are). Since the graphs are different, they make different assertions about which variables have causal effects on which other variables. If we can experiment, those claims can be checked directly. If a model says X is a parent of Y, but when we experimentally manipulate X it makes no difference to Y, we can throw that model out.

If we cannot experiment, we look for a qualitative, observational difference between the models — some conditional independence relation which one model says is present, and the other says is absent. For instance, in homework 10,

we had cancer$\perp\!\!\!\perp$tar|smoking in one model, but cancer $\not\perp\!\!\!\perp$tar|smoking in the other. To discriminate between these models, we just need to be able to test for conditional independence.

Recall from two lectures ago that conditional independence is equivalent to zero conditional information: $X \perp\!\!\!\perp Y|Z$ if and only if $I[X;Y|Z] = 0$. In principle, this solves the problem. In practice, estimating mutual information is non-trivial, and in particular the sample mutual information often has a very complicated distribution. You *could* always bootstrap it, but often something more tractable is desirable. Completely general conditional independence testing is actually an active area of research, though unfortunately much of the work is still quite mathematical (Sriperumbudur *et al.*, 2010).

If all the variables are discrete, one just has a big contingency table problem, and could use a G^2 or χ^2 test. If everything is linear and multivariate Gaussian, $X \perp\!\!\!\perp Y|Z$ is equivalent to zero partial correlation[1]. Nonlinearly, if $X \perp\!\!\!\perp Y|Z$, then $\mathbb{E}[Y|Z] = \mathbb{E}[Y|X,Z]$, so if smoothing Y on X and Z leads to different predictions than just smoothing Z, conditional independence fails. To reverse this, and go from $\mathbb{E}[Y|Z] = \mathbb{E}[Y|X,Z]$ to $X \perp\!\!\!\perp Y|Z$, requires the extra assumption that Y doesn't depend on X through its variance or any other moment. (This is weaker than the linear-and-Gaussian assumption, of course.)

The conditional independence relation $X \perp\!\!\!\perp Y|Z$ is fully equivalent to $\Pr(Y|X,Z) = \Pr(Y|Z)$. We could check this using non-parametric density estimation, though we would have to bootstrap the distribution of the test statistic. A more automatic, if slightly less rigorous, procedure comes from the idea mentioned in Lecture 6. If X is in fact useless for predicting Y given Z, then an adaptive bandwidth selection procedure (like cross-validation) should realize that giving any finite bandwidth to X just leads to over-fitting. The bandwidth given to X should tend to the maximum allowed, smoothing X away altogether. This argument can be made more formal, and made into the basis of a test (Hall *et al.*, 2004; Li and Racine, 2007).

Notice that this basic idea, of checking the conditional independence relations implied by a model, can be used even when we do not have two rival models. (This is more like a goodness-of-fit test than a comparative hypothesis test.) As usual, it is simple to reject a model whose predictions do not match the data. Managing to match the data is only evidence *for* a model if such a match was very unlikely, if the model is false. I will not, however, repeat the earlier discussion of the logic of model-checking here.

All of this is in fact fairly conventional hypothesis testing, where models are just handed to us by the Angel, or drawn out of scientific theories. The one wrinkle is that the DAG presents us with a lot of hypotheses which are in a sense small or local, making them easier to test, but which still bear on the global model. (We do not have to check a complete model of the determinants of cancer, just whether tar predicts cancer after controlling for smoking.) This is very suggestive. If we could paste together enough of these qualitative con-

[1] As you know, the partial correlation between X and Y given Z is the correlation between them, after linearly regressing both on Z. That is, it is the correlation of their residuals.

clusions about which varies are independent of which others, could we actually discover the right graph from the data?

2 Causal Discovery with Known Variables

Causal discovery is silly with just one variable, and too hard with just two for us.[2]

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.[3] 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);
- $X \leftarrow Y \rightarrow Z$ (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.[4] We can eliminate other edges based on similar sorts of background knowledge: men tend to be heavier than women,

[2] 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.

[3] 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.

[4] 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);

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.
- 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

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.

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.[5] 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 $X \rightarrow Y$ 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 taking advantage of any possible short-cuts. A refinement, called the PC algorithm, tries to minimize the number of conditional independence tests performed, essentially by doing easy tests first, and using what it can glean from them to cut down on the number of tests which will need to be done later (Spirtes *et al.*, 2001, §5.4.2, pp. 84–88). There has been a recent revival of statistical work on the PC algorithm, since the paper of Kalisch and Bühlmnann (2007), and at the very least it makes a good default procedure.

2.1 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.

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

2.2 Software

The PC and FCI algorithms are implemented in the stand-alone Java program `Tetrad` (`http://www.phil.cmu.edu/projects/tetrad/`). They are also implemented in the `pcalg` package on CRAN (Kalisch *et al.*, 2010, 2011). This package also includes functions for calculating the effects of interventions from fitted graphs. The documentation for the functions is somewhat confusing; see Kalisch *et al.* (2011) for a tutorial introduction.

2.3 On Conditional Independence Tests

The abstract algorithms for causal discovery assume the existence of consistent tests for conditional independence. The implementations known to me mostly 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), though the `pcalg` package does allow users to provide their own conditional independence tests as arguments. 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.

3 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,[6]

$$\lim_{n\to\infty} \Pr\left(\widehat{G}_n \neq G\right) = 0 \tag{1}$$

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 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. 1 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

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

make the required data size as large as he likes by weakening the dependence, without ever setting it to zero.[7]

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.

[7]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.

4 Exercises

To think through, not to hand in.

1. Describe how to use bandwidth selection as a conditional independence test.
2. When, exactly, does $\mathbb{E}[Y|X,Z] = \mathbb{E}[Y|Z]$ imply $Y \perp\!\!\!\perp X|Z$?
3. Would the SGS algorithm work on a non-causal, merely-probabilistic DAG? If so, in what sense is it a *causal* discovery algorithm? If not, why not?
4. Read Kalisch *et al.* (2011), install the `pcalg` paper, and give it the data from homework 10. Does it recover the graph which generated the data? If not, why do you think it failed?

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.

Hall, Peter, Jeff Racine and Qi Li (2004). "Cross-Validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association*, **99**: 1015–1026. URL http://www.ssc.wisc.edu/~bhansen/workshop/QiLi.pdf.

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.

Kalisch, Markus, Martin Mächler and Diego Colombo (2010). *pcalg: Estimation of CPDAG/PAG and causal inference using the IDA algorithm*. URL http://CRAN.R-project.org/package=pcalg. R package version 1.1-2.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H. Maathuis and Peter Bühlmnann (2011). "Causal Inference using GRaphical Models with the R Package pcalg." *Journal of Statistical Software*, **submitted**. URL ftp://ftp.stat.math.ethz.ch/Research-Reports/Other-Manuscripts/buhlmann/pcalg-software.pdf.

Li, Qi and Jeffrey Scott Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton, New Jersey: Princeton 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.

Sriperumbudur, Bharath K., Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf and Gert R.G. Lanckriet (2010). "Hilbert Space Embeddings and Metrics on Probability Measures." *Journal of Machine Learning Research*, **11**: 1517–1561. URL `http://jmlr.csail.mit.edu/papers/v11/sriperumbudur10a.html`.

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.