A Theory Of Inferred Causation

Summary of the second chapter of the book Causality by Judea Pearl
Causality Reading Group, 23. May 2006
Daniel Küttel, ETH Zürich

Introduction

The second chapter of Pearls book not only introduces us to a theory of inferred causation but also gives us some practical tools as well: Namely the two algorithms IC and IC*.

Those two are some quite generic algorithms asking only for an empirical joint distribution and then outputing some kind of causal model which has some rational behind it. (Mainly backed up by the theory introduced in the first part.)

Although he leaves out some details about the algorithms, for us it's the first encounter with some kind of general method to infer causation and ultimately infer a causal model/network.

The Setting

We want to find the cause-effect relationships (causal model)... and we are trying to do it in the hardest possible setting: We are not allowed to manipulate nature. The only thing we can do is observe nature (which may hide information). There are no controlled experiments.

The motivation why this should be possible is that humans do it all the time *and* usually agree (independently) on cause-effect relationships. The fact that we usually agree shows that this process can't be to random after all and has to have some mechanism behind it.

We have seen that it's difficult to get a grip on causality with standart statistical notions like simple statistical dependence. The implication works only in one direction. If there is causality you will see some statistical dependence but unfortunately not the other way around. (Statistical dependence is weaker than causality.)

So the main problem is always not to mistake spurious covariance for a causal relationship. But we also have to deal with things like hidden variables/information. (For example, if you have a *hidden* common cause for two effects, there is no cause-effect relationship between the two effects though it may look alike.)

Simple Observations

There are a couple of easy things that may give us some clues about the underlying causal mechanisms.

If we have temporal information at our disposal we can use it to „orient the edges“ because cause-effect relationships will never go backward in time. This still leaves us with finding possible causal connections but gives us the directions for free. For the most part of the chapter he won't be using temporal information.

Some statistical patterns (dependencies, independencies) can be associated with certain causal interpretations which are either the only possible or the only „natural“ explanations.

Search Space

We didn't spend much time on the definitions and I was just trying to give some intuition on them.

A causal structure D (def. 2.2.1) is just the DAG or the blueprint. The nodes and the edges are there but there is „no life“ in them. There is no functional or probabilistic dependence defined yet for the edges.

A latent structure L=<D,O> (def. 2.3.2) is an extension of a simple causal structure D. O is the set of observable nodes. We need this to model structures where nature hides some information from us.

A causal model <D,Θ> (def. 2.2.2) has a structure D (I believe it can be a causal or latent one) and a set of parameters Θ that assigns each node its dependence on its parents. This is the full model as we know it.

The introduced dependencies can be functional or probabilistic and they can have some random disturbances. Since we are going for a approximation of nature we may need those random disturbances to account for details we didn't model. This way we can control the complexity of the model: Just narrow down the parents of a given node if you consider it to

complex. This will introduce some random disturbance on that node because we removed some (hopefully unimportant) causes. As long as the random disturbance stay small it's a good trade-off between simplicity and correctness.

At what point did we simplify to much? In the chapter he states that the markovian condition (a node is conditionally independent of all its nondescendants given its parents) is worth protecting. For that the random disturbances have to be mutually independent and if we narrow down the parents of some nodes to much the „random disturbances“ won't be so random anymore. (Which kills the markovian condition. The random disturbances are the invisible connections.)

In our talk there was some discussion about the markovian condition and wether it's important or not. The thing is that you can just add some latent variables (capturing the dependence between those random disturbances) to restore the markovian condition.

The main question was: Should we still try to protect the markovian condition or can we rely on the latent variables?

Using latent variables the cool thing is that you can model the world at whatever granularity you wish and just let the latent variables do the rest.

Now that we know what kind of models we are looking for we just have to find the one wich is consistent with our empirical joint distribution (observed nature).

Unfortunately it's ambiguous and the search space is unbounded. (We have been there so no surprise here.)

Refitting the Search Space

To remove some ambiguity the basic idea is to always use the simplest working model. In order to identify a least complex model you have to come up with a „is simpler than“ order.

A latent structure L is simpler than a latent structure L' (def. 2.3.3) if L' can mimic L (only looking at the observables). In other words we are looking at the expressive power of a model. The less expressive power the simpler the model.

This order relation is certainly more semantic than a simple order based on number of nodes, number of edges or maybe number of layers in the model. But it's also more difficult to actually compare two models because in general you can't just tell which one is simpler from looking at the v-structures or something like that. Especially if you have hidden nodes.

It turns out that the thing that really kills expressive power in a causal model is not the existence of edges but the absence of which. This is obvious as soon as you realize that the independencies embodied in the structure restrict the expressive power and not the dependencies (because a dependency can „mimic“ an independency).

A minimal latent structure L is now a structure which has no other preferred structure (only looking at at structures that generate our empirical distribution).

We have been discussing about this notion of minimality. The main concern was that it doesn't take into account the complexity that may be in the dependencies we introduce. We haven't been quite clear wether you could possibly „trick“ the minimality: Just build a causal model which has „less nodes“ but extremely many states per node and extremely complex dependencies.

After some time we had to continue with the lecture but I think that you can't trick the minimality this way. I'd say that minimality is just a property of a structure independent of what dependency functions you might choose. (In that way minimality is taken over all the possible dependency functions you might want to introduce.)

The set of minimal solutions still has more than one candidate in it. That's why we can further restrict the set of solutions. The minimal solution should also be stable. Some independencies are embodied in the structure no matter what dependency functions you will choose (going from the causal structure to the causal model). But some independencies may be introduced using some special functions (going from the causal structure to the causal model). A stable structure (def 2.4.1) is now a structure which doesn't allow for such „numerical“ independencies.

Inferred Causation

After having „refitted“ the search space the solution to a given empirical distribution $P_{empirical}$ is a minimal latent structure L which is stable and

consistent: $P_O(L) = P_{empirical}$ (comparing the joint distribution over the observables)

Definition 2.3.6 says that whenever a variable C and a variable E have a cause-effect relationship in every minimal L consistent with $P_{empirical}$ we can infer causation from C to E.

In other words: If all explanations we can find think C causes E then C causes E. It seems as if you are trying to stay on the safe side. Whenever definition 2.3.6 infers causation it has to be real causation but it may be that there is causation and 2.3.6 stays silent just because there was one „bad“ minimal latent structure.

Some questions came up:

- Do such „bad“ minimal latent structures exist at all?
- Could one little approximation error in the empirical distribution prevent us from inferring a causation?
- Could we benefit from a less strict inferred causation? (Maybe have some probability on cause-effect relationship?)
- What if all minimal solutions let us infer causation but there is a bigger (not minimal) solution which can do it without causation for C and E? (Is this possible anyway?)

There wasn't much time to go into the remaining topics because the hour was nearly over. I will just briefly summarize them but there will be no more notes about the discussion because we pretty much had to rush through the remaining slides.

IC and IC*

Now that we can identify a unique (up to d-separation) solution we can formulate an algorithm to find it.

Both IC and IC* are quite similar: You start by trying to find a set of isolating variables for each pair of variables (making them independent). If you can't do so those two variables have to be closeley related and they get connected by an undirected edge. In the resulting undirected graph you just try to orient as many edges as you can subject to rules that should generate the correct orientations.

Those orientation rules are different for IC (working on causal structures) and IC* (working on latent causal structures). See page 84 and 88 for the details.

Working on latent structures (with hidden nodes) the search is again restricted because you could build arbitrarily complex hidden networks above the observables.

Instead of looking at all latent structures only special structures will be investigated: Projections (def. 2.6.1) of latent structures. A projection consists of it's observables and hidden variables which are always common causes of only two observables. (Resulting in only one „hidden layer“ above the observables.) Since in such a projection every pair of observables can have a hidden common cause it's possible to simulate/mimic any kind of dependencies (possibly generated by a more complex hidden model in nature).

Local Criteria for Causal Relations

Some statistical patterns allow us to infer causal relationships. Since we are only observing we can't do controlled experiments. Instead you can look at a third variable which does the experimenting for you. I like to call this „uncontrolled experiments“.

To find a genuine cause (def. 2.7.2) from X to Y you need a Z that controls X. If Z now can control Y as long as you don't use X to shield it you have found a genuine cause.

To find a spurious association (def. 2.7.3) between X and Y you have to find a Z that controls X without effecting Y and vice versa ruling out both directions of causality. This leaves only a (hidden) common cause as explanation.

Using temporal information makes those two tasks somewhat easier because you already now the direction/orientation of the possible cause-effect relationship.

Since we don't use temporal information in most techniques introduced here we can „infer time“ from our resulting causal model. Following the arrows is following the time and it's interesting (or encouraging) that usually this indeed coincides with the physical time.