Causality

Christina Heinze-Deml

1 Lecture 1

1.1 Different types of experiments

Can be viewed in a tree-like structure.

First differentiation is whether a study has controls, or not. Without controls, one can't say very much.

Next, if We do decide to go with controls, what kind of controls are they? They can be contemporary, i.e. happening now, or in the past, i.e. historical data. One has to be wary of confounders in historical data.

If it's contemporary, i.e. a study is being performed in the present, controls may be chosen. If they are not chosen at random, one once again needs to be very careful about confounders.

If the control is randomly chosen, one ought to do double blind trials, where neither the person receiving nor the doctor administering the treatment knows whether the treatment is true or a placebo. This leads us to randomized double-blind trials, the gold standard.

1.2 Internal v.s. External Validity

This is with respect to statements within studies.

For example, in the Salk vaccine trial, in the case where the trial group were people who consented to the vaccine, and the control was randomly drawn from that same group, it is internally correct to make statements about the trial - given that the findings are significant, they are significant for that group.

What may not hold is external validity - is the vaccine equally effective for groups that are different from the treatment group in the trial? External validity, to me, can be thought of as generalization error - do the statements that are true within the study generalize to the broader population etc.

1.3 Simpson's Paradox

Be careful when averaging over groups - sometimes it may be wise to do so, other times not so.

Essentially there may be lurking fixed effects between groups which would change the direction of the association. More on this later I suppose.

2 Lecture 2

2.1 DAG models

Parents of a node are immediate ancestors.

Children are immediate descents.

In general, a DAG decomposes probabilities such that

$$P(X_1, X_2, X_3 \dots X_n) = \prod_{i=0}^{n} P(X_i | X_{\text{parents}(i)})$$

Recall Markov chains - these have that old chestnut, the **Markov property**:

$$X_{i+1}, X_{i+2} \dots \perp\!\!\!\perp X_{i-1} \dots X_1 | X_i$$

In case of DAGs, We have a generalization of this is the **directed graph Markov Property**, which states that

$$X_i \perp\!\!\!\perp X_{\text{Non-descendants of } X_i} | X_{\text{Parents}(X_i)}$$

Since parents insulate the child.

Now, there's also the **Markov blanket**. The idea is that if You condition on the parents of a node, the children of a node, and the children's parents, then the node is independent of everything else no matter what. I think D-separation shows up here.

We need to think a little more about the descendants case, however.

2.2 *d*-Separation

So we have four types of connections (for illustration see page 407 of advanced data analysis from an elementary point of view):

Forward chain, backward chain, fork, collider

The first thing to observe is that *statistical* information flows both ways in a DAG. Observing a parent by definition gives information about the child, but observing the child also gives clues about the parent.

Now We think about the case where Z, the inbetween variable, is not known.

In the forward chain case, X gives information about Z, and knowing stuff about Z tells us information about Y, so $X \not\perp\!\!\!\perp Y$.

In the backward chain case, knowing X makes it's parent Z more likely, and knowing Z is more likely tells us stuff about Y.

In the fork case, where Z is a parent of both X, Y, knowing X still makes Z more likely, which makes Y more likely.

Finally, let's think about the collider.

I can see why knowing something about the collider explains away the other parents of the collider. This I buy.

Yeah I buy the black hole perspective - a thing has two causes, and We observe one and the thing maybe happened, maybe not. Therefore the two

causes are still independent. If We observe the cause and effect, this explains away the other cause.

Now for the case where the middle variable is known - chains block information flow by Markov property, as do forks, but colliders explain away.

So! We are interested in deciding whether variables X, Y are independent given some set of variables S. We say that X, Y are independent given S if every path from X to Y is blocked. A path is called **blocked** if it is not active. A path is **active** if every variable along the path is active. A variable Z is active if, as before:

Z is a collider and in S,

Z is a collider and a descendant of Z is in S,

Z is not a collider and not in Z.

All of which should make sense given the reasoning about chains, forks and colliders.

Conversely then, a variable Z is blocked if

Z is in S and not a collider,

Z is a collider and neither it nor any of it's descendants is in S.

if S blocks all paths between X, Y, then We say that X, Y are **d-separated** by S.

3 Lecture 3

3.1 Do-Calculus

So We have

$$p(y; \text{see}(X = \tilde{x}) = p(y|\tilde{x})$$

Which is just classical stats. To get do calculus We have

$$p(y; \text{do}(X = \tilde{x})$$

That there refers to the distribution of Y given that X has been forced to take a certain value - some sort of intervention has taken place.

3.2 Regime indicator

Is basically do calculus. We introduce some regime variable σ and do

$$p(y; \sigma = \text{whatever})$$

Where whatever is whatever You want. If it's the null symbol, You get back to standard stats. You can say regime a is where You forced X to be some value, and then $p(y; \sigma = a)$ You are talking about the distribution of Y given an intervention on X as specified by σ. The set of regime indicators is S.

Another important bit is what stays the same given different regimes. Distributions that don't change given different regimes are called *stable* or *invariant*.

3.3 Potential Outcomes

Just seem like regime indicators to me.

3.4 Causal Effects

Are formalized by

$$p(y; \sigma = a) \neq p(y; \sigma = b)$$

For different regimes a, b which change some X.
There is also **Average Causal Effect**:

$$E(Y; \mathrm{do}(X = \tilde{x})) - E(Y; \mathrm{do}(X = x'))$$

There are also individual casual effects and their interpretation is unclear to me. They are also cross-world, so. This is all cross-world actually.

3.5 Intervention Graphs

One way of going about it is just to add a regime node. The DAG is called an intervention DAG if the regime node is a source node (i.e. no incoming edges), the original DAG is a valid DAG (and We're just adding a regime node to the original valid DAG) and d-separation between regime node and any node in the DAG implies lack of causation - changes in regime node won't affect d-separated stuff.

Sidenote: **markov equivalence** of two DAGS is: suppose You have some DAG. Any old DAG. Now You can write down the conditional independence equations for that DAG, basically d-separation. Now switch the directions of the edges around a bit. If You switched the edge directions but the conditional independence equation is the same, then You have two DAGs which are markov equivalent.

3.6 Causal DAGs

Just change the interpretation of edges to be causal. So d-separated things are causally independent, and do-calculus forces probabilities to be one.

4 Lecture 4

Okay so none of the above was necessary lol.

Main bit about causal DAGs - the causal structure is assumed. So when You perform a "do" operation, You just swap out the value of that node in the DAG and replace the probabilities on that edge to 1 for the case You are interested in. Same for calculations.

So when writing down decomposed joint distribution, the probability of the parent vanishes and probability of child becomes the conditional distribution.

4.1 Simpson's paradox

How to pick a part when to condition on variables.

Well, recall that if You perform a do on a causal DAG, it's equivalent to removing incoming edges to that node.

So in the case of gender, We perform a do and now the graph has treatment $\rightarrow$ recovery and gender $\rightarrow$ recovery.

But, We want just treatment $\rightarrow$ recovery. So We condition based on gender to remove it.

In the case of blood pressure, there is no confounder, just treatment affect two variables. Performing a do still gives You all the information You need. I think.

5 Lecture 5

Okay so two variables share a confounder if there exists a path from confounder to both of them.

SEMs are just more nuanced ways to write DAGs. Parents become arguments of functions, children become stuff that depends on the output. Really just a causal DAG but each node gets a function that shows how the node depends on the parents.

Usually linear, in which case they are called linear structural equation models.

Clicker Q - true, false, true, true.

There are different kinds of interventions:

Surgical just sets a value

Shifting just adds noise.

Imperfect cuts incoming edges and replaces with some constant + noise.

5.1 Total Causal Effect

So there exists a total causal effect between X and Y if

Performing do X=x1 and X=x2 yields different distributions for Y.

Performing P(Y—do(X=x1)) $\neq$ P(Y)

5.2 Finding causal structure

Total average causal effect: look at all paths between two variables and add the weights. That the path method.

We can also **regress** to find the total average causal effect. To find effect of X1 on X4, regress with X4 as target and X1 as predictor, plus the parents of X1.

Adjustment set is the set such that regression on X plus adjustment set yields valid total causal effects.

6 Covariate Adjustment

6.1 Determining adjustment sets

There's a core definition here - any set Z that is a **valid adjustment set** must satisfy :

$$p(x_k|do(x_i)) = \int_{X_Z} p(x_k|x_i, x_Z)p(x_Z)dx_Z 0 \tag{1}$$

So basically, for any valid adjustment set Z, if We marginalize that set out, We get the do distribution.

Any set that satisfies that equation is called a valid adjustment set.

The valid adjustment set with the fewest elements is called the minimal adjustment set.

6.2 Backdoor criterion

So We're looking for a graphical way to decide on the adjustment set.

Recall We decided that We could just take the parents of a node before, but that turns out to be inefficient, or rather there's a subtler way of going about it.

So We have some nodes i, k and We're interested in $p(x_k|do(x_i))$.

Some adjustment set Z which does not contain i, k then satisfies the **backdoor criterion** if:

1. $Z \cap desc(i) = \emptyset$

2. Z blocks (d-separation) all "backdoor paths" from i to k in the graph. A path is a backdoor path if there's an edge entering i. So if the path includes (and it kind of has to end with) an edge into i, the thing We want to have a *do* distribution of, then that path must be blocked.

Okay so let's think about that for a second.

The first criterion can maybe be thought of as just avoiding collider bias.

The second criterion simply avoids confounders. This is identical to the reasoning in the "the parents of a node are a valid adjustment set" case, but now We just make the observation that it's not necessarily the parents We care about, it's actually the confounders.

Clicker question: yes, no, yes, yes, yes, yes, no.

6.3 Positivity

Okay so recall identifiability is about whether or not We can actually identify causal relationships - if there's a confounder for A and B but We cannot observe it, then good luck to us trying to make any kind of causal statements.

Okay so We perform a *do* on some node i. Positivity just means that the support of the distribution prior to the do has to be the same as the support after the do.

For exam type situations, just means that $p(x_i, X_Z) > 0 \; \forall x_i \in X_i, x_Z \in X_Z$.

6.4 Gaussians

So the point is that if everything involved is Gaussian, then just do linear regression with the target being k and the regressors being i and the adjustment set.

6.5 Adjustment Criterion - Shipster et al, Perkovic et al

According to those nerds, You just need these two conditions for a valid adjustment set:

1. Z does not contain any descendants of nodes $r \neq i$ on a directed path from i to k

2. Z blocks all paths between i and k that are not directed from i to k.

So the second criterion is basically same as before - don't want confounders.

Now what about that first one? Z cannot contain any descendants of nodes that are along the directed path from i to k. Suppose We did include a descendant, so what? Colliders suddenly being opened? Yep that's the one.

6.6 Efficient adjustment sets in linear SEMs

So We can do total causal effects via adjustment, but how good is that estimate? Turns out that can depend on the variables We choose.

Intuitively, We want to avoid stuff highly correlated with X_i, since then our features suck, and use variables that help predict X_k, since, You know, that'll give an actually good estimate.

So without further ado:

Let causal nodes $cn(i,k)$ be nodes $r \neq i$ on a directed path from i to k.

Let forbidden nodes be descendants of causal nodes and node i.

And then the optimal valid adjustment set is given by $pa(cn(i,k)) - forb(i,k)$.

Neat. The variance of the covariates produces by the optimal set will be less than or equal to the variance produced by any other set.

6.7 Direct Effects

As opposed to total effects. I don't think We need to care at all, other than the fact that in a linear SEM these are given by the weights.

7 Week 7

7.1 Identifiability

Again, an interventional (do) distribution is identifiable is You can calculate it just given the observational distribution and the graph structure.

We already saw that if We have an adjustment set, We can calculate this.

Now We look at two other means - frontdoor criterion and instrumental variables.

7.2 Frontdoor criterion

Suppose We know of an unobserved confounder. This means We're out of luck with adjustments - no way to remove the backdoor path to i.

However! The idea now is to instead of looking at i, We can look at all the nodes i influences, and how *those* influence k. We've sort of just moved a little bit farther in.

Another intuitive way of looking at it is that We're looking at the mechanisms by which i affects k.

Formally, a set M satisfies the frontdoor criterion with respect to i, k if

1. M blocks all directed paths from i to k.
2. There are no unblocked backdoor paths from i to M.
3. i blocks all backdoor paths from M to k.

Well the first condition makes sense - M accounts for all ways i could influence k.

The second requirement says: a path from i to M which is a backdoor path (so the edge is going *into* i must be blocked. What does this give us? Mechanisms g .oing from i to k but not back again?

The third requirement just ensures causality again - We don't want paths from M going to k through i since We're trying to get rid of i here.

7.3 Instrumental Variables

The idea is simple - use a variable that affects *only* i.

A variable is instrumental if

1. I affects i.
2. I is independent of confounders We can't observe.
3. I affects k only through.

Makes sense. There could be confounders but We get around them by using something that only affects the thing We causally care about.

7.4 Conditional Instrumental Variables

Okay so now We'll condition on some set S. The idea is that hey, I might affect k AND some other shit, but maybe We can just control for the other shit. The other shit is S.

Conditions We need:

1. I affects i given S.
2. Every path from I to k that is not blocked by S has an arrow pointing into i. The idea is that I affects k only through i when We control for S.
3. S does not contain descendants of i.

Seems to make sense. The last one is about not messing up effects of i on k.

7.5 Estimation of instrumental variables

With Two stage least squares.

Pretty simple. First We regress with i as target and S and I as predictors. Then replace i with the estimate, I think, and then preform another regression with the target being k and predictors being new i and S.

Clicker - yes and yes.

7.6 Identification strategies

Randomized controlled trials are self explanatory.

Covariate adjustment - but this only works if You already know the causal structure.

Frontdoor criterion - different way of adjusting I think.

Same deal for instrumental variables.

7.7 External Validity and Transportability

External validity is about whether the findings of a study are valid in external settings.

Transportability is about, I think, taking the identified causal effects and using them in a different setting.

Honestly not much to it - transporting with causal assumptions is fine, not fine if You're using proxies and the proxies differ in populations.

8 Week 8

8.1 Counterfactuals

Are simple. You observe one thing and ponder what if You observed another.

Assumptions:

SUTVA - stable unit treatment value assumption. Each patient is independent, each case independent etc.

Consistency - $Y = Y(x)$ for all cases, so the mechanism remains the same.

Ignorability - no confounders, and treatment assignment is independent of outcome.

Positivity - all cases occur, same as before.

8.2 Matching

Is one way to get around counterfactuals.

The simplest way is to take a positive case of treatment and match it with a non-positive case. If there are fewer treated than untreated people You'll lose data, and vice versa.

Matching is done via a metric of Your choice, e.g. Mahalanobis distance.

Can be done greedily or optimally, where optimal is feasible up to 1 million pairing, so 1k treated 1k controls.

Must also check that the treatment and control groups aren't too weird (no idea to be honest) with standardized mean difference. < 0.2 is fine, otherwise bad.

8.3 Propensity Scores

Summarize a whole lot of factors as just probability of getting treatment. Think of it as logistic regression. Can then match on propensity scores.

9 Week 10

9.1 Inverse Probability Weighting

If You have 9 untreated patients and 1 treated, one possibility is just to weight the treated patient 9x.

The rule is to weigh the treated samples by the inverse of their propensity score, and weigh controls by 1-propensity score.

9.2 Markov Properties

Global Markov - if A and B are d-separated by S then they are independent given S.

Local Markov - non-descendant independence given parents.

Markov factorization - probability of some node decomposes according to parents.

9.3 Minimal I-map

I-Map is a synonym for a causal DAG, and minimal I-map is just a minimal DAG.

To construct one just take the full factorization and simplify it as much as possible.

A skeleton is a DAG without the directions.

9.4 Faithfulness and perfect maps

So We had the global Markov property, which was that d-separation implied conditional independence.

Now We're going the other way around - a DAG is faithful if a conditional independence in the actual distribution implies d-separation in the DAG.

If a distribution if Markov and Faithful w.r.t. a DAG, then the DAG is a perfect map of the distribution.

There are distributions that a DAG can't capture - no perfect map exists.

9.5 Graph Terminology

A triple is a v-structure if You have $a \rightarrow b \leftarrow c$.

An **unshielded** triple, so I guess like a triple-candidate, is when a, b are adjacent, b, c are adjacent, but not a, c.

9.6 Markov Equivalence

A perfect map, if it exists, is unique up to Markov Equivalence.

Where, two DAGs are Markov Equivalent if they result in same d-separations.

And this is why We cared about v-structures - You can identify whether a v-structure exists by testing for conditional independence. You can't do this in a chain, since regardless of the direction of the chain, the same d-separation holds.

Therefore, DAGs in a markov equivalence class have the same skeleton with the same v-structures.

Note that the skeleton just gives You edges, but the v-structures actually give You directionality.

With this in mind, We have CPDAGs - completed partially directed acyclic graphs. These have undirected edges where the edge can be in either direction.

9.7 Constraint-based structure learning

The idea is simple:

1. Two nodes are adjacent if they cannot be d-separated.
2. If two nodes are not adjacent, then they can be d-separated.

So I suppose You test for stuff like that and v-structures to get a DAG up to markov equivalence.

10 Week 11

10.1 SGS algorithm

So You start with a full graph, which is to say an arbitrary ordering and then just expand the conditional probabilities.

So now for all pairs of nodes, asses whether they are independent given some other set of nodes, and if they are, remove the edge. Awful.

Now You need to check for v-structures. If two nodes are independent without conditioning but become dependent conditioning on an unshielded triple, then You have a v-structure.

10.2 PC algorithm

Now You look at adjacent nodes and trim their parents by considering parents of either node (just taking subsets of parents and checking for conditional in-

dependence). Speeds things up since now You aren't looking at just the whole graph.

Clicker - yes, yes.

11 Week 12

Establishing conditional independence is hard in non-Gaussian case.

11.1 Challenges for SGS and PC algorithms

Contradictory test results due to small sample size.

Choice of tuning α, the significance level for conditional independence.

Non-parametric conditional independence tests are hard.

11.2 Score-based causal structure learning

What it says on the tin.

First, You assume a parametric model - some set of parameters that can capture the entire search space.

Then come up with a loss function - in this case penalized log-likelihood, and optimize away. Can also use Bayesian Information Criterion.

BIC is neat because the score it gives is the same for all DAGs in the same Markov equivalence class (MEC), and it's decomposable into parts which means it can be efficiently computed.

12 Exercises

12.1 Series 1

Q1 a)

Evidence suggests it does but does not prove it - the problem is that the "examined" group of women opted in to the checks, and therefore they may be more health-oriented by default.

In other words, the treatment group is not randomized due to the choice involved, and so causal relationships can't be derived (though they may be suggested).

Q1 b)

Because the groups have identical make-up. Sure, the treatment group was asked to engage more actively in checkups for breast cancer, but this fact does not change their likelihood of dying from an illness other than breast cancer (or maybe changes the odds very slightly).

Q1 c)

Feels like a trick question - the refused group consists more of poorer women, and their rate of breast cancer is higher. The rate is lower amongst the examined group, which are supposed to be richer.

Ah the control group. So the idea is that taking just the poors gives You a rate of 1.5, but taking a mix gives You 2.0, so the poors are less affected.

Q1 d)

Since rich people suffer from fewer diseases, and the examined group consists primarily of the rich, this relationship between wealth and disease accounts for the disparity.

Q2 a)

Observational. Don't know if You could call it a study.

Q2 b)

Yes.

Q2 c)

Yes.

Q2 d)

Not without further interpretation.

One, if the gene is random, it should appear equally in controlling and non-controlling families.

But the presence of the obesity gene may make parents more controlling, so.

Q2 e)

Children's large weight makes parents more controlling in order to get the child to lose weight.

Q2 f)

Depends on the underlying causal structure. Overall, no.

Q3 a)

Non-descendants of D: A.

Q2 b)

Okay so to be separated from F We need to have C. But knowing C opens the collider ACD, so, can't be separated.

Q2 c)

Aren't A and D separated by default? Just don't mess with it.

Otherwise, C, F cannot be present, ever. Other than that, all nodes are game - G, B, E, H.

For the second path, We can do B and then the chain is blocked so E, H, G are all fine.

So either B and any subset of GBEH, or any subset of GH

Q4 a)

Suppose the opposite - two nodes are connected without the use of a collider. But since they share no ancestors, and there are no colliders, they also cannot share descendants, and so this contradicts our assumption that they are connected.

Q4 b)

Either the two nodes are disconnected and are therefore d-separated since no path exists.

Or they are connected by a collider, but neither the collider nor any descendents of the collider are known, so that's also fine.

12.2 Series 2

Q1 a)

Okay so the path method.

Paths from 2 to 4 are: $2 \to 3 \to 4$ and $2 \to 4$.

So the sum is just 3*5+4=19.

Q2 a)

C is standard gaussian, E is gaussian with mean 0 and variance 17.

Q2 b)

Okay We are not doing the conditioning lol.

Conditioning $E|C$ gives mean of eight and variance one.

$C|do(E = 2)$ leaves C untouched, and likewise the other way around.

Q3 a)

Hmmm, yeah?

If I intervene on X, in the first case Y should not change, in the second case it will. And then You write out the full factorization so $p(x, y) = p(x|y)p(y)$ in the one case and $p(x, y) = p(y|x)p(y)$ in the other, then do a *do* and see that they are not symmetrical.

Q4

12.3 Series 3

Q3 a)

Recall We had causal nodes that are nodes on a directed path from i to k, and forbidden nodes were descendants of causal nodes and the node i.

Uh so there is only one directed path, and it's the straight link between X and Y.

Parents of Y are A_2, B_2, R.

So that's the parents of the causal nodes I think.

Forbidden nodes are descendants of causal nodes and of node i.

So F is out.

Okay so it's just A_2, B_2, R.

Parental is trivial.

Minimal

Q4 a)

I mean, depends. Old people die anyhow, so.

Q4 b)

Older people die more.

I don't think this is a Simpson's paradox? Simpson's paradox can be described as individual groups displaying one trend, and overall groups displaying a different trend.

Q4 c)

At first glance it might, but upon closer examination it does not.

Perhaps white people are more susceptible, but if so then younger people would show more deaths.

12.4 Series 4

Q2 a)

Yep it's an instrumental variable.

Q2 b)

Well first We perform a regression with X as target and Z as predictor. We then replace values of X with the regressed/predicted values, and then perform a regression with the standard adjustment set (so with H included).

Q4 a)