Notes POC1

Artur Gaspar

May 21, 2024

Contents

Chapter 1

Causality Chapter 1: Introduction

1.1 Section 1.1: Introduction and review

1.1.1 Odds and likelyhood

Odds are the fraction of probabilities. **Prior (predictive/prospective) odds** is $\frac{p(H)}{p(\neg H)}$, and the **Posterior (diagnostic/retrospective) odds** is $\frac{p(H|e)}{p(\neg H|e)}$. This is how much more likely the hypothesis is to be true than false a priori and after observing the event e.

The **Likelyhood Ratio (Risk Ratio for epidemology)** is $\frac{p(e|H)}{p(e|\neg H)}$, remembering that Likelyhood is a function of B in $p(A|B)$, while the probability is a function of A.

The formula is: Posterior Odds = Prior Odds $\times$ Likelyhood Ratio.

My interpretation of $p(H|e)$ is the probability we give to H in the world where e happens, thus if we do $\frac{p(H|e)}{\neg H|e}$ we're seeing how more probable (multiplicatively) H is to be true in this world, and if we do $\frac{p(e|H)}{p(e|\neg H)}$ we're seeing how much more likely is the event e to happen in the world in which H is true than in the world in which it's not (it's a comparison accros worlds).

I interpret the likelyhood ratio as how many more times the evidence appears in the world where H is true than in the world where it's not.

So how more likely the hypothesis is to be true than false, after we observe the event = how more likely the hypothesis was to be true before the observation was made times $\times$ how many more times the evidence appears in the world where H is true than in the world where it's not.

Odds of hypothesis after e = odds before e $\times$ how much more e happens in H than in $\neg H$.

1.1.2 Coariance, Correlation, Regression Coefficient

Covariance is the expected value of $(X - E[X])(Y - E[Y])$, distance to the averages, $cov(X, X) = var(X) = (std(X))^2$, and **Correlation** is $corr(X, Y) = \frac{cov(X,Y)}{std(X)std(Y)}$.

Regression coefficient when estimating Y using X is $corr(X, Y) \times \frac{std(Y)}{std(X)}$, which is how much Y will change by unity of X we change, if we use the line that minimizes the quadratic error of the Y estimate. I kind of interpret this as $\frac{\text{(X-unities)}}{\text{(X-unities per standard devition of X)}} \times corr(X, Y) \times std(Y) =$

(number of standard deviations of X)$\times corr(X,Y) \times std(Y)$ = (number of standard deviations of Y)$\times$ (Y-unities per standard deviations of Y) = (Y-unities). The strange thing with this interpreatation is that $corr(X,Y) = (\frac{\text{standard deviations of X}}{\text{standard devations of Y}}) = \frac{\text{standard deviations of Y}}{\text{standard deviations of X}}$, is the function that given one ammount of standard deviations returns the other one... Maybe this is a reflection of the limitations of the linearity assumption?

1.1.3 Axioms

Finally, the graphoid axioms for independence of random variables (all of them conditioned on Z, and I simplified a little bit):

1. **Symmetry**: X is independent of Y iff Y is independent of X
2. **Decomposition, Weak Union and Contraction**: X is independent of YW iff ((X is independent of Y) and (X is independent of W conditional on Y)). This is not how it's written in the book, but I think this single affirmation is equivalent to the Decomposition, Weak Union and Contraction axioms.
3. **Intersection** (only for strictly positive distributions): X is independent of W given Y and X independent of Y given W implies X independent of YW.

Summary:

1. Independence is symmetric.
2. Being independent from two things is equivalent to being independent to one alone and the other given the first one. In other words, being independent from two things means that looking at the value of one does'nt help and looking at the other after knowing the first doesn't help as well.
3. Being independent from two things (if nothing is impossible (?), maybe so we can condition on anything?) is the same as being independent from the first even if you know the second and being independent from the second even if you know the first.

If X is independent of Y, then $p(x|y,z) = p(x|z)$, we can ignore the irrelevant information.

1.1.4 Counter example for third axiom

The third axiom does not hold for instance in the following joint distribution (A, B, C are the random variables with two values each):

1. $p(a1, b1, c2) = \frac{1}{2}$.
2. $p(a2, b2, c1) = \frac{1}{2}$.
3. All other probabilities equal 0.

Then we have $p(a1|b1,c2) = 1 = p(a1|b1) = p(a1|c2) \neq p(a1) = \frac{1}{2}$ and $p(a2|b2,c1) = 1 = p(a2|b2) = p(a2|c1) \neq p(a2) = \frac{1}{2}$.

It doesn't make sense to talk about other conditional probabilities, as they are conditioned on something inexistent. We can say that A is independent of B given C, and A is independent of C given B, but A is not independent of BC.

This happens here because some values of BC are impossible, so we kind of know the value of C only by knowing B and vice-versa... So we know C iff we know B, and then after we learn one we don't need the other, but we can't ignore both.

If anything was possible, we would have $p(a1|c1) = p(a1|b1,c1) = p(a1|b1) = p(a1|b1,c2) = p(a1|c2) = k$, so $p(a1) = p(a1|c1)p(c1) + p(a1|c2)p(c2) = k(p(c1) + p(c2)) = k$, so the axiom follows.

1.1.5 Why conditionals are enough to specify independence

Just a disclaimer: $p(b1,c1) = 0 \rightarrow 0 = p(a,b1|c1) = p(a|c1) \times p(b1|c1) = p(a|c1) \times 0$, so to satisfy the independence we really just need to specify it for possible "worlds" (the values k that make $|k$ possible)...

If we write the independence with the "and" way, we get $p(a1,b1|c2) = 1 = p(a1|c1) \times p(b1|c2) = p(a1,c2|b1) = 1 = p(a1|b1) \times p(c2|b1)$, and the same for the other one, but it seems more complicated to me, the only advantage would be that we could write the zero parts, $0 = p(a1,b2|c2) = p(a1|c2) \times p(b2|c2) = 1 \times 0$. Let's try to use only the conditional version.

1.2 Section 1.2: Bayesian Networks

1.2.1 1.2.1) Conventions

skeleton of a graph is the undirected version of it.

In this book, a **path** might not follow the direction of the edges.

Family of a graph is a node and it's parents.

Root is a node without parents and **sink** a node without children.

Tree is a connected graph with at most one parent per node (one node can point to many but only one node can point to it), and **chain** is one with at most one child per node (onde node can point to only another one, but many can point to it).

1.2.2 1.2.2) Bayesian Networks

One of the main goals is to represent an joint distribution with less data, which is possible if every variable is independent of almost all others.

The **Markovian parents** of a node is a minimal set of nodes that, conditioned on them, the value of the node is independent from the value of all other nodes. It's a set of variables that we can condition on to ignore the rest when estimating the initial node, but such that we can't remove any variable from this set.

This set is unique if the joint distribution is strictly positive, and this implies an unique Bayesian Network.

I believe that if it's not strictly positive, then if we consider that the parents must be minimal, it might be impossible to draw a Bayesian Network, otherwise we accept non-minimal sets of parents

and acknoledge taht we might have more than one BN. See the subsection "Instability of parents for non-positive distributions".

We say that G represents P, or G is compatible with P if we can decompose P with the information we extract from G (the DAG). For instance, $p(a1, b1, c1, d1, e1) = p(b1|a1)p(c1|a1)p(d1|b1, c1)p(e1|d1)$ is the decomposition for the graph with $A \to B$, $A \to C$, $B \to D$, $C \to D$ and $D \to E$.

1.2.3 Instability of parents for non-positive distributions

I think that it's possible to have more than one minimal set (of Markovian Parents) if the third graphoid axiom is not satisfied, because then we can have X independent of Y given Z and of Z given Y, but not on YZ. So we might want to require the distributions to be strictly positive...

Take for instance the following joint for A, B, C binary:

1. $p(a1, b1, c2) = \frac{1}{2}$.
2. $p(a2, b2, c1) = \frac{1}{2}$.
3. All other probabilities equal 0.

Here if we know one value we know the other two, so $\{B\}$ or $\{C\}$ are minimal markovial parents for A; $\{A\}$ or $\{B\}$ are minimal for C; and $\{A\}$ or $\{C\}$ are minimal for B. So, we kind of can't create an undirected graph that represents the dependencies well... One node will connect to other two, but it acually depends on only one (any one)...

1.2.4 1.2.3) d-separation

This is a criterion to extract the conditional independences between variables from the graph.

X, Y and Z here can be sets of more than one variable.

We say that a *path* is **d-separated** or **blocked** if either Z has a variable in the middle of the way or as a confounder, or the path has a collider which is not in Z and no descendent of the collider is in Z.

We say that Z d-separates X from Y if it does so for every path from X to Y.

It's really important to consider the descendent part! Conditioning on a variable unblocks every collider that is reachable in reverse order (following the arrows reversed) from this variable.

X and Y are d-separated by Z if and only if for all distributions compatible with the independencies of G, X and Y are conditionally independent given Z. Also, if they are not d-separated, almost all distributions make they dependent (they don't say "how much indepedent").

Selection bias, **Berskon's paradorx** or **explaining away effect** is the situation in which after conditioning on one variable we render two others dependent (knowing that one does not have a specific value lets us increase the chance of another, for instance).

Observational Equivalence is the situation in which we have two graphs such that any distribution compatible with one is also compatible with the other.

It happens iff they have the same undirected structure and the same "v-structures", which are converging arrows without a connection between their tails: $X \to Y \leftarrow Z$ but no arrow between X and Z forms a v-structure.

1.2.5 1.2.4) Inference with BNs

The book comments a bit on how we could try to estimate conditional probabilities of some variables given the observation of others. I'm not going to focus on this.

1.3 Section 1.3: BNs with causal directions

A Causal Bayesian Network is a Bayesian Network with causal directions.

We say that a distribution after an intervention is compatible with the CBN if it's Markov relative to it (we can decompose the joint with respect to the BN, or the parents make the childs independent of non-descendents), the chance of the interventions happening is one, *and the conditional probabilities remain the same for variables we didin't act on.*

The joint after the intervention can be factorized as $P(v) = \Pi_{i|V_i \notin X} P(v_i|pa_i)$, which is basically the original joint without the $P(v_i|pa_i)$ of the variables we acted on. v is a vector here (the entrances are the values of the random variables that are represented by the nodes of the CBN).

Two properties: $P_{pa_i}(v_i) = P(v_i|pa_i)$ = interventions are according to the conditionals, and $P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$ = no interventions besides the one in the parents can influence a variable

Pearl argues that the advantage of causal models is to transport results to other environments and predict the results of changes that aren't purely observacional.

1.4 Section 1.4: Counterfactuals

1.4.1 Laplacian vs Stochastic model

The Laplacian one has deterministic functions and unobserverd probablistic variables, he stochastic one is more similar the Bayesian Network approach, if I understood correctly

Pearl says that this is more general than probablistic functions, but to me this just makes sense if by stochastic he doesn't mean something like a Markov Chain instead of the function, as this would certanly be more general... The BNs do not really have Markov Chains, but conditional probabilities, maybe that's what he means?

1.4.2 1.4.1: Structural Equations

Structured Equation Models are defined defining each variable as a function of the parents and unobserved variables (erros). If it's linear, then its a **Linear Structured Equation Model**.

One important point: Pearl says that it's possible to estimate counterfactuals with data and a causal model, and to test empirically whether they hold or not. I believe he will focus on how to do that in later chapters.

In the linear models, the coefficients are the variation rates per forced variation of a value, in the sense that it's how much the value would change if we changed only that value by one unity.

It's usually assumed that the error terms are independent, if they are dependent we represent a dotted double-headed arrow between the variables involved.

The hyerarchy of Causal problems defined by Pearl are:

1. **Predictions** are the "what if we found out that the value of this other variable was this?"
2. **Interventions** are the "what if we set the value of this other variable to this?"

3. **Counterfactuals** are the "what would be the value of this variable if the value of this other one was that instead of this?"

1.4.3 1.4.2: Probabilistc Predictions (and Definitions and equivalences between SCMs and BNs)

Causal Diagram is the diagram obtained by connecting the parents to the childs according to the structural equations. If this graph is a DAG, then it's **semi-Markovian**, and if the erros are independent, then it's **Markovian**. If it's semi-markovian, the joint is completely determined by the distribution on errors.

If the model is markovian, then this is a valid Causal Diagram: given the parents, a node is idependent of all other non-descendants. The proof is just to get the full graph, with the errors, then notice that we can remove the errors without losing independencies.

Pearl says that this is implied if we include every variable that might be a causa of two or more others, and that there is no correlation without causation...

The idea seems to look at the data and determine all probabilities first, even without knowing the deterministic functions (and the errors or distribution on errors) themselves... For any joint distribution compatible with a bayesian network, there is always at least one Functional Model with this same network (and Pearl mentions that usually there are infinitely many) that generates it with some values for the error/unobserved variables.

So, I think this is what he meant before, that the functional models are more general: we can encode in them anything we could encode in a BN.

1.4.4 1.4.3: Interventions

Four advantages mentioned by Pearl of using the graphical representation of Causal Models are:

1. The conditional independencies do not depend on the specific functions themselves, so if we can represent something in the causal model even with limited information, and given the model we don't need to compute anything to know whether some variables are independent given others (this is also possible with BNs, isn't it? We can also just build the graph without the probabilities and check independencies).

2. It's simpler to specify the connections, and the model has few parameters (I would argue that BNs has the same number or less parameters, the advantage to me is actually that the functions are finite, while the probability distributions are not, but then the distributions on unknowns are also infinite).

3. It's simpler to think of whether or not the parent set has all relevant variables that are a direct cause of some variable, instead of checking whether they make this variable independent of the others when we condition on them (and are a maximum set that does that). (we kind of could do this for BNs, right? But yeah, we would need to think that the independence is guaranteed, I think I agree with this one)

4. If something changes, the change might be local on some variables only, and with these models we can model this change by changing less the model, instead of recomputing everything from scratch. (This really does seem like a big advantage, if we change from one country to another the functions will change, and the conditional probabilities of the BNs change,

but the functions might be simpler. Again, the unknowns might change as well but I don't doubt at all that it's simpler to determine the unknowns than to re-estimate the conditional probabilities)

1.4.5 1.4.4: Counterfactuals

The idea is to say which variables were responsible for some result. For instance, if someone takes an experimental treatment to a disease and dies, did they die *because*, *despite* or *regardless* of the treatment?

Pearl says that we can treat counterfactuals as, instead of what would have happenned with X_1 if $X_2 = y$ instead of $X_2 = x$, what will happen if we reverse the outcome and repeat the experiment keeping everything equal except the value of X_2. This is called the *persistency* assumption.

I didn't understand why the assumption is necessary, and how exactly can we reach the conclusion for the assumption, but Pearl says that the proportion of people that died and recovered are equal with or without taking treatment, then (ignoring sampling variances) the proportion of people that died under treatment but woudn't if not treated would be the same than the proportion of people that didn't die without treatment but would under treatment. The idea seams to be that if the treatment is $x\%$ rensposible for the death of someone, then it would be $x\%$ responsible for the death of someone alive and untreated; if $x\%$ of the treated dead died because of the treatment, then $x\%$ of the alive untreated would have died if treated.

Two different situations given as examples that generate the above data but have different counterfactuals are: the treatment has no effect or half of the population has an allergy that protects them from the disease but kills them if they receive treatment. In the first case, treating someone untreated wouldn't change anything (all of the dead untreated would still be dead if treated), in the second group everyone that died under treatment was allergic and would still be alive if untreated.

The basic idea, viewing the SCM as a CBN with the unknowns explicited, is to Bayesanly-update the values of the unobserved variables given the observations, then intervene with the alternative values of whatever we want to know the alternative, then re-compute the distribution after the intervention. Viewing as Structured Equations, I think that we set the values of the observations to estimate the values of u, then set the new values of the alternative world and recompute everything. Pearl divides this into the following:

1. Abduction: basically estimate $P(u)$ from the observations.
2. Action: basically do the intervention, "bend the course of history minimally to comply with the hypothetical condition".
3. Prediction: Compute the desired probability.

Pearl says it's possible to compute estimatives without the full functinons between nodes and without knowing the distribution of unknowns, with just some assumptions of both.

1.4.6 Questions and confusions

I still am a bit confused about being able to have more than one set of parents per node if the distribution is not strictly positive... What do we do about that? What if there is a logical limitation, and an example that's better (and harder to find the problem) than just two equal

variables causing another? Would everything break or is it stable to lead to an "almost zero" probability when it would be zero?

I didn't get why the assumption of $p(y|x) = \frac{1}{2}$ of (1.46) was necessary for the exercise "left for the reader". I think I will be able to do this later.

1.5 Section 1.5: Some terminology

Apparently, *probabilistic* stuff are quantities obtained from the joint, and *statistical* stuff is obtained from the joint of observables, ignoring non-observables completely.

Causal stuff are things defined in terms of a causal model.

Chapter 2

Chapter 2: Theory of Inferred Causation

2.1 Section 2.1: Introduction and intuitions

Basically, we're going to use the basic structures (for instance, $X \rightarrow Y \leftarrow Z$) and their statistical results to try to infer the causal relationships. The ideia is to get causal directions that are likely, not necessarely certain (like in inference in general).

2.2 Section 2.2: A framework for causal discovery

We're going to assume that the reality is that everything is deterministic but some things are unknown, and that we have a DAG, which represents the structure of what causes what.

He re-defines the causal model here, which is kind of a function BN with uncertainty in the (independent) unknown variables.

Pearl mentions that we could (conceptually) start with an arbitrarily well detailed causal structure to represent the universe, and then generalize it by aggregating variables untill we can't generalize anymore without losing the properties we want to keep. He argues that one such property is the Markov condition: to keep the errors independent.

He argues that we intuitively think of correlations without a common cause as spurious, and that we consider "strange" to have them. So, our models should have this property to better reflect our intuitions.

We can then leave some causes to be summarized as probabilities (in the unknowns), but not if they also affect other variables.

Latent Variables are defined as unknowns that affect more than one variable in the system.

The idea is basically that we ask questions about the probability distribution of some set of observable variables and try to infer the (hidden) causal model of reality from it.

2.3 Section 2.3: Occam's Razor

In the scenario in which all variables are observed, X has a causal influence on Y if there is always a directed path from X to Y in every minimal structure consistent with the data...

Latent Structure is a causal structure on V with some $O \subseteq V$ observables.

Θ_D are the **parameters** of the causal model D! (The parameters are: the distribution on the independent disturbancies u_i and the deterministic functions from parents and disturbancies to the values).

One latent structure L is **prefered** (smaller in the semi-partial-order relation) to another L' if and only if for any parameters of L we can find parameters on L' that mimic the results of the distribution of observed variables. They are equivalent iff one is prefered to the other and the other to the one.

So, the prefered is the simpler, the semi-order relation is kind of a "complexity" notion. This definition is in terms of *expressivity* results, it's not defined in terms of number of parameters or anything "synthatic". So, we would prefer one model to another even if the first one has few free parameters.

If there are no hidden variables, then (as I understood it) two networks are equivalent iff they lead to the same conditional independencies.

Now we define that X **has a causal influence on** Y if there is always a directed path from X to Y in every minimal latent structure (from the set of available ones) consistent with the distribution we observed.

My impression is that if we don't have have any unblocked paths between two variables, then they must be independent. If we have unblocked paths, then they might be dependent (but if we have two paths, for instance, they might cancel out each other).

Pearl says that sometimes patterns in the distribution unambiguously implies a causal relation (by assuming minimality only), making no assumption at all about the presence or absense of latent variables.

2.4 Section 2.4: Stable Distributions

If A and B are random coin tosses, and C is the XOR between them, then any two variables are marginally independent but dependent conditional on the third. Pearl says that any of the three configurations with a collider would be valid, and are indistinguishable by only looking at the data.

To avoid this kind of thing (avoid parameters that lead to these problems, in this case the XOR computation), he imposes a restriction on the valid distributions compatible with some given model, which he called **stability**:

A causal model *generates a stable distribution* if and only if we do not lose any independency of this distribution no matter how we set the parameters of the model (we can't get less independencies by changing the parameters).

The XOR example of the coins do not generate a stable distribution, because if we changed the function or the hidden distributions, the independencies might have changed (if the first coin is more likely to be heads, then).

Numerically:

$$p(result = 1|first = 0) = 1\%(chanceofsecondcoin = 1)$$

$$p(first = 0) = 90\%$$
$$p(result = 1|first = 1) = 99\%(chanceofsecondcoin = 0)$$
$$p(first = 1) = 10\%$$

In this case, the result is not independent of the outcome of the first coin, as:

$$p(result = 1) = (1*90+99*10 = 90+990 = 1080)/100^2 = 0.1080\% \neq 1\% = p(result = 1|first = 1)$$

In other words, a distribution is stable if the independencies obtained there are the independencies we can identify via the graph of the model! *If the distribution has more independencies than those visible in the model, then it's unstable!* It can not have less because the model implies its independencies (for compatible distributions, obviously).

Perl says like this stability is about small changes, but the definition itself allows any change, no matter how small. Is it possible to have a situation in which a small change of parameters does not delete some of the conditional independencies, but a big change does? It depends on what is called small here obviously, but I'm under the impression that no, this should not happen very often...

He calls the stable independencies *structural independencies*, that do not depend on the specific numeric values. And it seems he just assumes that this really reflects small changes (particularly, I'm under the impression that some specific equalities must hold for instable independencies to hold)...

2.5 Section 2.5: Recovering DAG Structures

Assuming that the distribution is stable, then there's an unique minimal causal structure up to d-separation.

patterns are the causal structures with undirected edges whenever some minimal causal structure has one direction and other minimal structure has the other direction for that edge.

The algorithm that builds the pattern given a stable distribution consists simply of creating the undirected graph with an edge iff no conditioning makes the variable independent, then add arrows as mediators iff a and b are not adjacent and c is a common neighbour that is not in the set that separates them. After that, orient any edges that if oriented differently would make a cycle or a new v-structure (remembering, v-structures are mediator structures).

Basically, it seems to me we're finding the undirected graph, directing it whenever we can see a collider, then directing the edges that are always directed that way if we don't have a directed cycle and we don't have unindentified colliders.

In summary, I think we can see the colliders only, we assume there's no cycle, and we can't differentiate between confounders and mediators.

Also, most importantly, this assumes we have the exact distribution, not an empirical observation of it.

Pearl mentions many ways of doing what the algorithm requires (how to do some parts is unespecified). There are fast solutions for linear gaussian simplifications.

I'll not note the details here, but some ideas are given in the book of how to imeplement the unspecified steps of the algorithm.

Pearl says that latent structures require special treatment, which will be covered a bit more in the next chapter. So, I guess all of this was for graphs without latent variables? Maybe he was

building just the "causal Bayesian Network" model... **It might be important to understand this better**.

2.6 Section 2.6: Recovering Latent Structures

A *projection* of a latent structure L is another latent structure L_O (O are the observables) that keeps all independencies from L (for every stable distribution generated by L) and in which each unobserved has two children (we omit the unobserveds and just connnect the two children with an undirected edge).

I didn't understand exactly what the first few phrases of this section mean...

The distributions are not unlikely anymore to have extra independencies? (maybe this is just saying that the unobserved variables create/remove some independencies, so the observed ones might not even have a DAG structure)

Also, why does he say that we can "no longer guarantee" a DAG structure among the minimal compatible latent structures? Before this point, we already could find a case in which we didn't find directions for all edges, right? We just found the "maximally directed" graph...

Also, why does the instability of the distribution imply that we we can't find a DAG structure???

Maybe he just means that it might really be the case that there might be no DAG consistent with the data?? (and minimal)? Not having stability imples that we can not guarantee that the independencies represent stuff in the graph, and therefore we might find only non-DAG minimal structures???? Something like we have one variable that acts as a mediator of some nodes and as a confounder of parents of these nodes, somehow???? So we can not assign a direction? This is super weird, I'm not sure that's what the book meant...

I also didn't understand very well how the projection helped, as they use stable distributions (are the stable distributions in the definition of projection not stable on O but stable on all variables?)...

Are we trying to find a projection of the complete model in respect to the observables, then trying to find a causal link on one projection on the minimal model implies finding it on every minimal model???? And what is a distinguished projection?

Also, I think that why the algorithm works is really not obvious, as Pearl says the original proof that it worked.

In the end, what I got is: we have an algorithm that identifies as much as it's possible to, the directions of the edges and whether we have an unobservable confounding, effectively buiding all we can gather from a distribution. Assuming that the distribution is somewhat stable (in respect to the whole structure...), that no small changes of parameters would change independencies.

This difference between stability with respect to the observables and with respect to the whole structure is still strange to me, but I'll assume that the stability with respect to the observables means just that we can't act as if the unobservables weren't there, we need to consider the possibility of latent variables interferring. If we do, we would be assuming that there is no extra independency besides the ones induced by the arrows between the observables, but some of these independencies might not be even present in the usual distribution generated by the complete model, and maybe (who knows), some might appear (two dependent variables are seem as independent, we would need to be able to make two dependent variables independent by adding edges and nodes, this seems impossible because the DAG-independency is just "there is no unblocked path", so if there is a path we would need to remove it otherwise we cant make the variables un-independent!).

2.7 Section 2.7: Local criteria

The IC* algorithm, which returns what we can infer about the causal structure from the data (assuming all latent variables are independent, I believe), can be used to define the different types of inferred causations we can get from data. All are based on using a "virtual controll" variable.

I believe we're assuming that the unobserved variables are not caused by the observed ones, that distributions are always stable according to the real model, that the unobserved are independent and finally I think we're rellying that we can represent any SCM with latent variable as a projection of it (each unobserved connects exatly two others and stabile original distributions can lead to observable-independence-equivalent stable distributions on the projection (I don't know why we need "exatly two" and not "at least two" here, and also why we need non-adjacent, this seems somewhat weird))...

X is a **potential cause** of Y if they are dependent in any context, and we have a third variable Z that influences Y without influencing X in some context.

Here **context** means some set of variables to condition on, and "influences" here means dependency, not being independent. This is the argument "the wet grass does not cause the rain because we can easily get the grass wet in a way that does not influence the rain at all".

What I get from a potential cause is: either X and Y are both caused by an unobserved confounder, or X causes Y (or some kind of combination of both). I think that the reason why other configurations are not allowed is because of that projection assumption, we're getting the projection of what's really happenning...

I believe that under stability and the "projection assumption", X and Y are correlated in any context if and only if they are both caused by an unobserved confounder or one causes the other. The "partial cause" simply removes the possibility of Y causing X.

X is a **genuine cause** of Y if they are dependent in any context, and there is a potential cause of X (let's call it Z) that influences Y but if we condition on X, it doesn't, under some context. If my last paragraph is true, then this is because we either have $X \to Y$, $X \leftarrow Y$ or $X \leftarrow L \to Y$ (dependent under any context), but the confounder is not possible because if it was when we condition on X we would connect Z to L using X as a collider, and then to Y.

But why can't we have $X \leftarrow Y$? Same thing! If we conditioned on X it would act as a (in this case, direct) collider between Z and Y, so Z would affect Y.

This means that genuine causes are potential causes, genuine causation is stronger.

One quick note: we actually consider the transitive closure of the above to be the "genuine cause" relation. Maybe the above could be called a "genuine direct cause"???

The condition of always being correlated in any context is callend **adjacency**.

The way to notice genuine causation is then to find something that is not caused by X but is adjacent, that changes with Y in general, but not if we condition on X. So, if X is rain and Y is wet grass, we can say that in any situation they are correlated, and we could use Z as the current season, for instance... This is not so intuitive to me, but mathematically seems sound.

X and Y are **spuriously associated** if they are dependent *in some context*, and there is a context and another variable that is correlated with X without being correlated to Y undder this context, and another context and variable that is correlated with Y but not X under this context.

Temporal information simplifies things, as we assume we can't cause something in the past:

X is a genuine cause of Y if there is a context and another variable such that both that come before X and the variable is related to Y, but if conditioned on X it's not, under this context.

X is spuriously associated with Y if they are dependent on some context, X comes before Y, and there is a variable that is independent of Y but dependent of X.

An **intransitive triplet** is a trio of variables a, b, c such that a and b are independent but a, c and b, c are dependent. Then, c can not cause a and can not cause b. That's because we can use b as a virtual control to check if c causes a, for instance: c and a are dependent, but knowing b affects c without affecting a.

2.7.1 Information Flow? And quick notes

In the end, all statistical things here are kind of about the knowledge we have on some variables... The uncertanties of their values, represented by the probability distribution.

It's interesting that knowing the confounder, no information about one variable X helps knowing the other Y, so the confounder Z is kind of the "maximum" you can learn about Y by looking at X.

This seems like there is some interesting relation between this kind of causal model and the quantification of information flow...

Also, it seems like we're assuming here that we have perfect distinction of independence between variables, but I assume that independence is somewhat unstable when sampling...

são coisas que por alto são intuitivas mas exigem uma necessidade gigante de formalismo enorme, porque tem vários corner cases, várias coisinhas que podem dar erradas, me lembra mais análise que álgebra essa matemática aqui

One last note is that all this formalism in causality seems somewhat like Analysis: we can have a lot of corner cases, and a lot of strange things hapenning, so all the formalism is necessary so we do everything correctly.

2.8 Section 2.8: Nontemporal causation

This looks more like a philosophy/curiosity chapter, so I'll not try to understand everything here completely.

Pearl is basically talking here about how strange would it be if something satisfied our definitions of causation but went in the reverse direction in time (something now causing something in the past, or something in the future causing something now).

He defines **statistical time** as any topological ordering of the variables according to at least one minimal causal structure consistent with the empirical distribution.

From what I understood, a Markov Chain has one statistical model that always goes foward, one that always goes backward, and others that given a fixed node (I believe we can fix a day for instance), we go foward and backwards from that node.

He says something about two coupled Markov Chains having only one possible time, conjectures that usually statistical time should coincide with physical time, and someone related this to the second law of the thermodynamics, but Pearl says it might be something else as the example does not follow that law (?) I didn't get this completely.

Also, as I understood it, it seems like we can change the direction of this unique statistical time sometimes by simply changing the coordinate system... He then argues that it's more a matter of how we prefer to represent things than something about the nature of reality. He also speculates that this kind of reasoning might have been naturally selected to give more value to finding out

what will happen in the future given present information than what happened in the past that explains the present...

By quicly googling it, it seems like Pearl is a physicist (curiosity because he presented some Physics examples here, that at least to me are a bit obscure).

2.9 Section 2.9: Conclusions

2.9.1 How to infer causation and assumptions

"No causes in – No causes out; Occam's razor in – Some causes out."

We use "virtual interventional experiments" to remove some possibilities of causal directions.

We also a kind of mediator strategy to check for genuine causes (X causes Y if I know of something Z that changes with X and is not caused by it: so I kind of can simulate an intervention on a cause of X, be it Z or the latent cause between X and Z. Such that if I change a cause of X then Y changes, but if I fix X and change the cause of X then Y does not change.

We can use temporal knowledge to simplify the conditions to infer causal relationships.

We are assuming both that the best causal model is the least general one (Occan's Razor), and that the observed distributions do not have accidental independencies (independencies that mathematically satisfy the concept of independency, but in general would be dependent)

Also, **we might or might not assuming that the joint distribution is strictly positive**, at least for the bayesian networks we needed this, here I'm not sure. We kind of could get ambiguous Bayesian Networks with non-strictly-positive distributions, and if we get such a causal BN an turn it into a causal model with latent variables, maybe the distribution is not stable anymore? Or the result from the IC^* just encodes the possible models in that "not sure if this is the parent, the child or the latent variable" mode...

2.9.2 Markov assumption

Pearl says that the Markov assumption (independency of distinct latent variables) was challenged, and that by enforcing it (then relaxing it by allowing latent structures), we're losing the ability to represent some models (models that can not be represented as a causal model with stochastic errors and a subset of observables). Pearl says it might not be a very big loss because we would not be able to do much with such types of model (he says "it's not clear" how we could do that).

Pearl says that only in the Quantum Mechanics world we have observations that can not be explained by latent variables, and it would be a "scientific miracle" if someone managed to replicate this kind of phenomenon in the macroscopic world.

2.9.3 Stability assumption

He says that equalities on paramters of SCMs (to make unstable independencies) lead to joint distributions with zero lebesgue measure, so they should never happen. Pearl says it's different from equalities on things like correlation coefficients because we want *autonomy* of the mechanisms of the model, we could vary them independently (the model parameters), so equality constraints that restrict them are counter to this idea and will not usually happen in natural conditions. He also argues that usually people will give stable examples when asked.

So, some people argued that the algorithms developed for this type of causal discovery (based only on correlations) are better if we have longitudinal studies conducted under varying conditions, so we can be more sure that the result we got is in fact stable. Pearl argues that even if it's true, being able to use this kind of reasoning is still better than relying only on controlled randomized trials.

2.9.4 Other things

Pearl says that causality has a different approach from traditional Machine Learning, as we're not done after fitting perfectly the data, and getting perfectly the joint distribution, we'll still have many possible valid causal structures and might need to do interventional experimentation or observe some virtual intervention.

2.9.5 Notes and doubts

One thing I think I finally understood: the SCM is the model with stochastic error terms and deterministic stuff. The **latent structures** are *NOT* correspondent to these error terms, they are the subset of the SCM that we can't observe! This is why we might have one latent variable (which might be an error term or not) pointing to two observable variables! Also, I believe that the projection just says that there is a way to represent an entire SCM in a simpler way, at least from the point of view of the observable part of the SCM and assuming consistency regarding the independencies of stable distributions of both models... I still didn't get though why we need exactly two and why nonadjacent variables in the definition of projection...

What if we don't know anything else besides the joint distribution of XY?

If X causes Y should knowing the value of X reveal more about Y than if Y causes X? Should I be more certain about the value of Y once I know X than I am of the value of X once I know Y?????

If Y is caused by another things, then I don't think this necessarily follows. If Y is caused only by X, then this seems plausible...

Doubt:

In Judea Pearl Causality, why does the definition of a projection need every unobservable to be a common cause of exactly two (instead of at most two) nonadjacent (instead of any two) observable variables????

Chapter 3

Chapter 3: Causal Diagrams and the Identification of Causal Effects

3.1 Section 3.1: Introduction

The focus of this chapter is going to be to formalize qualitative assumptions about the causal structure and their consequences. The goal is to derive causal inferences from the combination of *data*, *qualitative assumptions* and *experiments*.

We'll be able to determine if we have enough information to infer the results of an intervention, if we do determine these results and if not determine what experiments and observations we might need to get that.

3.2 Section 3.2: Intervention in Markovian models

3.2.1 3.2.1: Recalling the concept of (Semi-)Markovian models and interventions

A full specified causal model is represented by the functions $x_i = f_i(pa_i, u_i)$ and a distribution $P(u)$ on the independent random disturbances.

DAGs are **Semi-Markovian** and independent disturbances makes the model **Markovian**.

The focus of this chapter is on Markovian and Semi-Markovian models. Non-Markovian models are more the focus of Chapter 7.

Pearl says this is the "nonparametric form" of the Structural Equations model, and if I understood correctly this is just because here we won't specify the functions themselves...

Recalling that we can represent an entrance of the joint as a product of each variable conditional to parents (in particular, if they were all independent, it would be the product of all of them):

$$P(x_1, x_2, ...x_n) = \prod_i P(x_i|pa_i)$$

Pearl defines the Causal Effect $P(y|do(x))$ as the probability of y obtained by removing the equations that define x from the model and setting $X = x$ in the other equations. In the graphical

version, it's obtained by removing edges going to X and setting $X = x$. The joint computation, I believe, is going to simply have 0 value for when $X! = x$ and otherwise $P(x_1, ..., x_n|do(X = x)) = \prod_{i:X_i!=X} P(x_i|pa_i)$, as the variables in X are set to x with probability one and so their conditionals would also be one (besides, we kind of disconnected their parents from them).

3.2.2 3.2.2: Interventions as variables

We can augment the causal model by adding a parent F_i to each node X_i that determines a distribution on functions (now these are Markov Chains, right?) and we can use this to represent different types of changes on functional relations besides just the constant "set to X_i value x_i". $x_i = I(pa_i, u_i, f_i) = f_i(pa_i, u_i)$. We can create a value $F_i =$ idle that represents no intervention, for instance.

We can use this to somehow view sudden changes in the functional relations as interventions (for instance, a tax reform in some economic setting).

3.2.3 3.2.3: Computing the effect of interventions

As stated before, when we set $do(X_i = x_i')$ we can just rewrite the joint in the conditional on paretnts form just by removing the conditional $P(x_i|pa_i)$.

He then says to multiply and divide by $P(x_i'|pa_i)$ (THE pa_i VALUES HERE NEED TO BE THE SAME AS THEY APPEAR IN $x_1, x_2, ..., x_n$), and we get (if $x_i = x_i'$, otherwise it's zero):

$$P(x_1, ..., x_n|do(X_i = x_i')) = \frac{P(x_1, ...x_n)}{P(x_i'|pa_i)}$$

Which implies:

$$P(x_1, ..., x_n|do(X_i = x_i')) = P(x_1, ...x_n|x_i', pa_i)P(pa_i)$$

Writing with X as the set of all variables not in $PA_i \cup Y \cup \{X_i\}$:

$$\begin{aligned}
P(x_1, x_2, ..., x_i', ...x_n | do(X_i = x_i')) = & \\
&= \prod_{j \neq i} P(x_j | pa_j) \\
&= \frac{P(x_i' | pa_i)}{P(x_i' | pa_i)} \prod_{j \neq i} P(x_j | pa_j) \\
&= \frac{1}{P(x_i' | pa_i)} \prod_{j} P(x_j | pa_j) \\
&= \frac{1}{P(x_i' | pa_i)} P(x_1, x_2, ..., x_i', ..., x_n) \\
&= \frac{P(pa_i)}{P(x_i', pa_i)} P(x_1, x_2, ..., x_i', ..., x_n) \\
&= \frac{P(pa_i)}{P(x_i', pa_i)} P(x, y, pa_i, x_i') \\
&= P(pa_i) P(x, y | x_i', pa_i)
\end{aligned}$$

So, we can get:

$$\begin{aligned}
P(y | do(X_i = x_i')) &= \sum_{x_1, x_2, ..., x_n : \text{neither vals of} X_i \text{ nor } Y} P(x_1, x_2, ..., x_i', ...x_n | do(X_i = x_i')) \\
&= \sum_{x, pa_i : \text{vals of } X \text{ and } PA_i} P(pa_i) P(x, y | x_i', pa_i) \\
&= \sum_{pa_i : \text{vals of } PA_i} P(pa_i) P(y | x_i', pa_i)
\end{aligned}$$

This is the theorem 3.2.2 of the book, **adjustment for Direct Causes**.

Intuitively, we can compute the result of the intervention $do(X_i = x_i')$ on Y by adding for every possible value of the parents of X_i the expression $P(y|x_i', pa_i)p(pa_i)$. Note that if PA_i have no influence over Y that's not through X_i, then this is basically $P(y|x_i')$. Anyway, we're averaging the effect of X_i on Y according to our prior knowledge on the possible values of PA_i.

If we want to do different types of intervention: if the mechanism that defines X_i changed, and now it uses parents PA_i^*, then the resulting joint distribution could be obtained by replacing $P(x_i|pa_i)$ by $P^*(x_i|pa_i^*)$, getting $P^*(x_1, ..., x_n) = P(x_1, ..., x_n)\frac{P^*(x_i|pa_i^*)}{P(x_i|pa_i)}$ (we just removed $P(x_i|pa_i)$ and added $P^*(x_i|pa_i^*)$).

I believe that in this part we're assuming strictly positive distributions, otherwise we might not want to divide and multiply by $P(x_i'|pa_i)$

Pearl gives an example of a Dynamic Process Control, which I won't get into detail here.

This allows us to determine the effect of interventions when all parents of where we are intervening are observable, now we will see what to do when they are not.

3.2.4 3.2.4: Identification of Causal Quantities

A quantity is defined as **identifiabile** in a class of models iff the quantity is equal for distinct models only if their generated distributions are also equal. If the observations are limited, then the quantity is **identifiable** from the observations iff the quantity is equal when the observations are equal.

In summary: we can't identify a quantity if it can have two different values for the same observation we can make.

Under the class of models that have the same graph (but maybe distinct f_i) and generate positive distributions on the observables, we define that the causal effect from X to Y is **identifiable** from a graph if the quantity $p(y|do(x))$ can be uniquely computed from any positive probability of the observables (the causal effects are equal for models generating positive probabilities and with the same graph).

Pearl says that positive distributions are necessary because we would not be able to infer $P(y|do(x))$ if $X = x$ never happens in the observed data. He says it's possible to extend this, but he won't now.

So, to show that we can't identify the causal effect, we just need to show two sets of values of f_i that induce the same positive distributions but with different causal effects.

3.2.5 Summary

The causal effect of X on Y is identifiable whenever we can observe X, Y, PA_X.

3.3 Section 3.3: Controlling confounding bias

From now on, I'll try to go faster, just quickly looking at what's written but not reading everything in detail.

This is about whether to condition or not on some variables.

The **back-door criteria** between X and Y is satisfiable by Z if no node in Z is a descendent of a node in X and all paths between X and Y with an arrow pointing to X are blocked. As we have a DAG, this path can't be from Y to X if there is a path from X to Y.

I like the intuition for the second condition of not having a way for the information to "flow" from X and Y from another source.

The intuition behind the first condition (**do not condition on descendants of** X) is that if we do that, then we're allowing colliders in the chain that goes up to X. If the last arrow is $K \to X$, then information on the error of K, ϵ_k, is going to flow to X. If K is in the way from X to Y, we're also letting information flow from the error to Y. So we actually need only to exclude descendents of X that are also descendents of some variable on a path between X and Y. *Excluding descendents is stronger than necessary!*

Back-Door Adjustment is the theorem that we can compute the result of *do* if we have a set of variables that satisfy the back-door criteria $p(y|do(x)) = \sum_z P(y|x,z)P(z)$.

Front-Door Criterion is when Z blocks all directed paths from X to Y, $P(Z|do(X)) = P(Z|X)$ (all paths from X to Z are free of unblocked back-doors) and all X blocks all paths from Z to Y. This way, we can write $P(y|do(x)) = \sum_z P(z|do(x))P(y|do(z)) = \sum_z P(z|do(x)) \sum_{x'} P(y|x',z)P(z)$. We need to get rid of confounders between X and Z and between Z and Y.

To use the front door criteria, it's required to have $P(x, z) > 0$ always.

The way I see it, *back-door* requires us to find all counfounding paths and *front-door* requires us to find all mediators between X and Y.

We can combine front and back-door, to get alternatives if necessary.

3.4 Section 3.4: A calculus of intervention

This is about algorithms on how to compute the results of intervention from available data:

The do-calculus.

One quick notation note: $P(x|do(y), z) = \frac{P(x,z|do(y))}{P(z|do(y))}$.

There are some strange rules for deleting observations, *do*s and turning *do*s into observations. I'm not going to extend myself into this as it just seems like the idea of the *do* formalized in a way that's usefull. Seems too formal, maybe it'll get simpler later.

We can infer the effect of some *do* if we can apply the rules Pearl specifies to remove all *do*s and end up with things we can observe.

These rules are sufficient for deriving all identifiable causal effects.

Sometimes, it's also possible to identify the causal effects from X to Y by conducting experiments on other variables on Z. This is called **surrogate experiments** and we can represent this by keeping the *do* operator only on variables of Z. For instance, we might want to control the diet to estimate the effect of cholesterol levels on blood pressure.

3.5 Section 3.5: Graphical tests of indentifiability

If we have a unobservable bow between X and Y, and no variable mediating, only the direct influence of X on Y, then we can't estimate the effect of interventions on X, not even if this is in a bigger graph.

This happens for instance, if the clinical experiment of a drug is randomized but who accepts is not.

Removing edges and adding mediator can never worsen the identifiability of causal effects.

Pearl shows examples of models in which the effect of X on Y is or is not identifiable and discusses them. I'll not focus much on that, but it seems like it's possible to learn how to compute $P(y|do(x))$ or if it's not possible by using the information here. If I understood correctly, there is currently no mechanical method of computing this automatically.

The idea is to identify interventional distributions.

Seems like this paper of 2013 shows exactly if it's possible to estimate the causal effects of interventions: "An Information Theoretic Measure of Judea Pearl's Identifiability and Causal Influence".

Also, local identifiability is neither necessary nor sufficient for global identifiability.

I passed very quickly through this, but it specifies how to estimate effects of interventions given a graph.

3.6 Section 3.6: Discussion

Comments on extensions of the causality ideas presented.

The applications of this language of causal graphs is expected to require a lot of domain knowledge.

Pearl mentions about how to translate from the theory of graph notations he presented to the potential outcomes theory, and relations to G-estimation.

There is a theorem at the end that says that it's possible to estimate the causal effect $P(y|do(x))$ if there's no bi-directed path (a path of bi-directed unobserved arcs) between X and any of it's children (not successors, *children*), and this is more general than any results of the chapter. (*so why not edit the chapter such that only this is presented?*). So, this is a sufficient condition! Note that the front-door criteria example is not a counter example, in that case only Z is a child of X!

This condition is *necessary and sufficient* if Y includes all variables except X. There are other sufficient and necessary conditions in another (2006) paper mentioned there, for the general identifiability of $P(a|do(b), c)$.

There's a roadmap of the results of the chapter as well: backdoor criterion, do-calculus, and the formal meaning of counterfactuals (I believe it's simply to apply the observations, estimate the unobservables u, then apply the "counterfactual intervention" and infer whatever we wanted to infer after that).