Causal Statistical Decision Theory|Why subjunctive probability?

David Johnston

July 23, 2020

1 Interventions may fail to be well-defined

Interventional causal models run into serious difficulties when some variables are defined to be non-invertible functions of other variables. This work sharpens long-standing criticism of interventional models which held that interventions are ill-defined. Hernán and Taubman (2008) claimed that "the causal effect of obesity on all-cause mortality" is ill-defined, as there are multiple actions that might alter someone's BMI, such as diet changes, exercise or gastric bypass surgery, each of these may have a different impact on someone's risk of death and that the notion of an "intervention on obesity" must somehow depend on what can actually be done to change someone's body weight. Cartwright (2001), along similar lines, points out that for many systems it may be impossible to perturb *only* the prospective cause leading her to doubt about how well defined "the result of perturbing only the prospective cause" actually is.

These prior arguments draw heavily on our intuitive ideas of what constitutes an intervetion. While our informal causal knowledge must be a foundational part of the interventional account of causality, it could be argued that the appropriate role of informal knowledge is to specify the causal mechanisms that underwrite the definition of the interventions, and it is inappropriate to use informal knowledge directly to define interventions. For example, Pearl (2009) writes that the worries over the possibility of perturbing only the prospective cause are an error caused by giving the actions an investigator can actually take undue priority:

> Thus, for Cartwright, a set of equations that share parameters is inherently nonmodular; changing one equation means modifying at least one of its parameters, and if this parameter appears in some other equation, it must change as well, in violation of modularity.
>
> Heckman (2005, p. 44) makes similar claims: Putting a constraint on one equation places a restriction on the entire set of internal variables. Shutting down one equation might also affect the parameters of the other equations in the system and violate the requirements of parameter stability.

> Such fears and warnings are illusory. Surgery, and the whole semantics and calculus built around it, does not assume that in the physical world we have the technology to incisively modify the mechanism behind each structural equation while leaving all others unaltered. Symbolic modularity does not assume physical modularity. Surgery is a symbolic operation which makes no claims about the physical means available to the experimenter, or about invisible connections that might exist between the mechanisms involved.

It is not clear to me whether informal knowledge of causal mechanisms really underwrites interventions, or if in fact interventions are supposed to underwrite causal knowledge. Nonetheless, it is at least plausible that criticisms of interventions that employ informal definitions of intervention may fail if an appropriate definition of intervention were adopted. I show here that under extremely mild restrictions of the idea of "intervention", causal effects frequently fail to be unique. The broad outline of our argument follows:

- For many variables X, it is true by definition that $\mathsf{X} = f(\mathsf{Z})$ for some non-invertible $f : Z \to X$
- Any purported "effect of X" is therefore actually some mixture of "effects Z" and due to the non-invertibility of f it is not clear which mixture ought to be chosen

Our argument was anticipated by Pearl (2018) himself when he responded to Hernán's criticism of "the causal effect of obesity as measured by body mass index" by substituting "the effect of BMI" for "the effect of the vector of many factors that describe obesity" (note that a function from a vector to a scalar is usually non-invertible):

> That BMI is merely a coarse proxy of obesity is well taken; obesity should ideally be described by a vector of many factors, some are easy to measure and others are not. But accessibility to measurement has no bearing on whether the effect of that vector of factors on morbidity is "well defined" or whether the condition of consistency is violated when we fail to specify the interventions used to regulate those factors.

I affirm that any difficulty in measuring the "underlying vector of factors" does not undermine the idea that this underlying vector might have some causal effect on morbidity. However, Pearl has notably not defended the idea that *BMI* has a causal effect on morbidity. Rather, he asserts that the effect of BMI is actually a composite of effects of an "underlying vector". I argue that *many* variable are vulnerable to having their "causal effects" collapse in the same manner.

Concretely, Shahar (2009) argues that because BMI is defined as a person's weight divided by their height, it is appropriate to say BMI is caused by a person's height and weight and nothing else. He argues further that once these

are included in a causal graph, BMI has no causal effects leftover - any possible "effect" of BMI is really just some combination of the joint effects of weight or height. Extending this argument, consider that a person's weight is by definition equal to the weight of the fat in their body plus the weight of everything else in their body. Therefore, any possible "effect" of a person's weight is really just some combination of the joint effects of the weight of fat in their body and the weight of everything else in their body. The weight of all the fat in a person's body is itself the sum of the weight of all the white fat, the weight of all the brown fat and the weight of all the beige fat in their body. Therefore, perhaps the notion that the weight of fat in a person's body has some causal effect is just an illusion, and what is actually under discussion is a combination of the joint effects of the weight of the brown fat in their body, the weight of the white fat in their body and the weight of the beige fat in their body. It's not clear that we ever arrive at something that supports a "true" causal effect, and if we do we clearly have a great deal of backtracking still to do. It is not at all clear how the enormous model that arises from all of this backtracking supports any approximate causal conclusion we could draw from a practical model that features variables we can feasibly measure (I expand on this below).

It is possible to define causal effects in CSDT that do not fail in this manner. This is because CSDT, unlike interventional models, does not guarantee that "the effect of X on Y" exists for arbitrary X and Y.

1.1 The main argument

Assume the following:

1. In order to determine the causal relationship between two variables X and Y we require a probability measure $P \in \Delta(\mathcal{E})$ and a set of background causal relationships $\mathbf{R}$ such that X, Y are random variables in $\mathcal{E}$ and $\mathbf{R}$ is sufficiently large to include all causal relationships relevant to determining the effect of X on Y

2. The causal relationship between X and Y is a function $P(\mathsf{Y}|do(\mathsf{X})) : X \to \Delta(\mathcal{Y})$

3. If Z is any variable that is on a backdoor path between X and Y in $\mathbf{R}$, then there exists a "joint interventional map" $P(\mathsf{X}, \mathsf{Y}, \mathsf{Z}|do(\mathsf{X})) : X \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$ such that the marginal on Y of $P(\mathsf{Y}, \mathsf{Z}|do(\mathsf{X}))$ is $P(\mathsf{Y}|do(\mathsf{X}))$, the marginal on Z, $P(\mathsf{Z}|do(\mathsf{X}))$, is $x \mapsto P(\mathsf{Z})$ for all $x \in X$ and the marginal on X, $P(\mathsf{X}|do(\mathsf{X}))$, is $x \mapsto \delta_x$ for all $x \in X$

4. If it is a basic definition of a given problem that a variable X is some function $f : Z \to X$ of the variable Z, then Z is a causal ancestor of X in $\mathbf{R}$

5. If it is a basic definition of a given problem that $\mathsf{X} = f(\mathsf{Z})$, then $P(\mathsf{Z} \in f^{-1}(x)|do(\mathsf{X} = x)) = 1$ for all $x \in X$. That is, the interventional map is such that $\mathsf{X} = f(\mathsf{Z})$ with probability one.

Suppose X takes values in $\{0,1\}$, it is a basic definition of a given problem that $\mathsf{X} = f(\mathsf{Z})$, and there is a causal path in $\mathbf{R}$ from Z to Y that does not contain X. By (4), Z is on a backdoor path from X to Y. Then, by (3) there is a function $P(\mathsf{X},\mathsf{Y},\mathsf{Z}|do(\mathsf{X}))$ such that $P(\mathsf{Z}|do(\mathsf{X}=0)) = P(\mathsf{Z}|do(\mathsf{X}=1))$. By (5), $P(\mathsf{Z} \in f^{-1}(0)|do(\mathsf{X}=0)) = 1$. Noting that $f^{-1}(0)$ is disjoint from $f^{-1}(1)$, we have $P(\mathsf{Z} \in f^{-1}(0)|do(\mathsf{X}=1)) = 0$, contradicting (3).

Assumptions (1) and (2) are universally endorsed by proponents of interventional causal models (Spirtes et al., 2000; Pearl, 2009; Woodward, 2016). Assumption (3) is strictly weaker than Pearl (2009)'s definition of the do-operator, though we also investigate a weaker verson of (3) below. Assumptions (4) and (5) is new, require further discussion.

In defense of assumption (5), suppose that an intervention could change the *definition of BMI*. Then the causal effect of BMI on mortality would have nothing to do with the effect of BMI *as defined as weight/height* on mortality. For this reason, we think it is reasonable to disregard causal effects that contradict basic definitions.

Assumption (4) is somewhat tricky, as it depends on the condition of assumption (1) that $\mathbf{R}$ is "sufficiently large". We cannot make this condition precise. Nonetheless, we offer two separate arguments that the notion of "sufficient largeness" should lead us to include a causal path from Z to X in $\mathbf{R}$ if it is a basic definition that $\mathsf{X} = f(\mathsf{Z})$:

- X is determined from Z by an autonomous mechanism: $\mathsf{X} = f(\mathsf{Z})$ for all values of all other variables, which is an autonomous mechanism, and autonomous mechanisms should be included in $\mathbb{R}$ as causal relationships
- Reasoning from "possible interventions": If f is not a single valued function, then there are two of values z_1, z_2 of Z such that changing from z_1 to z_2 compels a change in the value of X, and this restriction applies to any "possible intervention". Thus an intervention on Z produces a change in X, and so Z is a cause of X

Note that these arguments are not equivalent. The first does not compel us to accept that X is also a cause of Z, while an argument analogous to the second does.

At this point we seem to have: if a variable X is by definition equal to some function of another variable Z and that variable may also be a cause of some third variable Y, then the "causal effect" of X on Y *cannot exist*. This seems to be quite alarming given that it is extremely common that we require some variable to be equal to a function of other variables.

There is a possible way around this contradiction: if we allow that X is also a cause of Z (which means that $\mathbf{R}$ contains a cycle), then we can allow interventions on X to alter the distribution of Z; this can be accomplished by weakening assumption (3):

3′. If Z is any variable that is on a backdoor path between X and Y in $\mathbf{R}$, then there exists a "joint interventional map" $P(\mathsf{X},\mathsf{Y},\mathsf{Z}|do(\mathsf{X})) : X \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$

such that the marginal on Y of $P(\mathsf{Y},\mathsf{Z}|do(\mathsf{X}))$ is $P(\mathsf{Y}|do(\mathsf{X}))$, and the marginal on X, $P(\mathsf{X}|do(\mathsf{X}))$ is $x \mapsto \delta_x$ for all $x \in X$

This avoids the previous contradiction - there is now nothing wrong with the support of $P(\mathsf{Z}|do(\mathsf{X}=0)$ being disjoint from the support of $P(\mathsf{Z}|do(\mathsf{X}=1))$.

The assumptions made are far too weak to uniquely define the interventional map $P(\mathsf{X},\mathsf{Y},\mathsf{Z}|do(\mathsf{X}=x))$. For example, all that we require of $P(\mathsf{Z}|do(\mathsf{X}))$ is (5), and there are many functions that meet this requirement. The theory of causal Bayesian networks does provide a unique definition of an interventional map, but this rule implies (3) which we have found to be too strong for our purposes.

We can propose a modified backdoor adjustment rule that substitutes $P(\mathsf{Z}|do(\mathsf{X}=x))$ for $P(\mathsf{Z})$, but this fails to be a satisfactory rule. If we make the additional supposition suppose that Z is the *only* cause of X in $\mathbb{R}$(X is, after all, uniquely determined by Z) then we have no backdoor paths between X and Z and no unblocked backdoor paths between X and Y after conditioning on Z. It follows that

$$P(\mathsf{Z}|do(\mathsf{X})) = P(\mathsf{Z}|\mathsf{X}) \tag{1}$$

$$P(\mathsf{Y}|do(\mathsf{X})=x) = \sum_z P(\mathsf{Y}|\mathsf{X}=x,\mathsf{Z}=z)P(\mathsf{Z}|do(\mathsf{X})=x) \tag{2}$$

$$= \sum_z P(\mathsf{Y}|\mathsf{X}=x,\mathsf{Z}=z)P(\mathsf{Z}|\mathsf{X}) \tag{3}$$

$$= P(\mathsf{Y}|\mathsf{X}) \tag{4}$$

Equation 4 holds for the causal effect of X on *any* Y. This is clearly unsatisfactory. We can turn instead to the theory of cyclic causal graphs presented in Forré and Mooij (2018). In this theory, the causal arrow from X to Z must be witnessed by another function $g : X \times U \to Z$ such that $\mathsf{Z} = g(\mathsf{X},\mathsf{U})$where U is a "noise" variable with some distribution $P(\mathsf{U})$ that is fixed under the intervention $do(\mathsf{X})$. In this case, it is easy to see that $P(\mathsf{Z}|\mathsf{X}=x) = g(x,\cdot)_\# P(\mathsf{U}) = P(\mathsf{Z}|do(\mathsf{X}=x))$ i.e. the interventional map of Z is again the same as the probability conditional on X. Furthermore, we also have some $h : X \times Z \times U \to Y$ such that $\mathsf{Y} = h(\mathsf{X},\mathsf{Z},\mathsf{U})$. It is clear that $P(\mathsf{Y}|do(\mathsf{X}=x),\mathsf{Z}=z) = h(x,z,\cdot)_\# P(\mathsf{U}) = P(\mathsf{Y}|\mathsf{X}=x,\mathsf{Z}=z)$. These two facts again imply 4!

Pearl (2017) has explored an "imaging" operator based on the intervention operator $do(\mathsf{X}=x)$ that is able to evaluate disjunctive "interventions". We might suppose that $do(\mathsf{X}=x)$ is equivalent to $do(\mathsf{Z} \in f^{-1}(x))$, in which case we might be able to make use of the imaging operator to evaluate $do(\mathsf{X}=x)$. However, as Pearl shows, the derived imaging operator implies that $P(\mathsf{Y}|do(\mathsf{Z} \in f^{-1}(x))) = P(\mathsf{Y}|\mathsf{Z} \in f^{-1}(x)) = P(\mathsf{Y}|\mathsf{X}=x)$.

1.2 Resolution?

The difficulties raised here require some elaboration of interventional models. Three possibile elaborations are:

- If a variable X is defined to be equal to $f(\mathsf{Y})$ where f is non-invertible, is it inappropriate to say Y causes X? If so, why?
- Are there some sets of variables that are forbidden from co-occurring in interventional models? If so, which sets are forbidden, and how can we be sure a chosen set of variables is acceptable?
- Are causal effects usually non-unique? If so, how should the non-uniqueness be handled?

2 Criticism of the potential outcomes system

Rough outline:

- Overall, it's under-specified and confused
- Some call consistency a definition, some an assumption. Note that consistency is "half of a similarity metric" a la Lewis, but the other half is nowhere to be found
- Some talk about interventions (see above)
- Some say "what would happen if you did an ideal intervention" - but is this a definition or an example?
-

The potential outcomes system of causal inference is under-specified. When someone judges ignorability to hold in a particular randomised experiment and judges it to fail for some other experiment, the potential outcomes system does not provide any more basic claims that constitute the assumptions. Rather, they are judged to hold or fall by direct appeals to intuition. Appeals to intuition render it difficult to direct pointed criticism at the theory, as the truth and falsehood of certain propositions depends on the intiutions of the people involved in the argument. This makes it a poor system for the collaborative pursuit of truth, as criticism plays an essential role in this.

It is in principle possible to ground the potential outcomes system in more fundamental assumptions. For example, Lewis (1986) has proposed that the "truth" of counterfactual propositions should be evaluated in terms of their truth values in the "most similar worlds" that make these propositions true. If it were possible to define a satisfactory measure of world similarity then assumptions like ignorability *would* reduce to more basic claims, and could be disproved by showing that a more similar world exists in which ignorability fails. As it stands, however, I am not aware of any suitable metrics of world similarity, and as I will show it is likely that no single metric will serve.

The potential outcomes system does feature the assumption of *consistency*. This is the assumption that the potential outcome Y^a "the value of Y on the supposition that a occurs" is equal to Y if a actually occurs. This precisely

matches Lewis' view that if a actually happens then the most similar world in which a occurs is the real world. So far so good, but the whole point of the similarity measure is to say something nontrivial about Y^a if a *does not* occur. The potential outcomes system is silent on this, handwaving the issue away with the admonition that this must be determined by expert judgement. They do not explain how one should obtain expertise in offering true answers to unanswerable questions.

Hernán and Taubman (2008): "Suppose that, in the observational study in the neighboring country, the data analyst compared the mortality of subjects who happened to have a BMI of 30 ($A = 1$) and a BMI of 20 ($A = 0$) at baseline. Now consider a study subject who had a BMI of 20 at baseline. It is not obvious that, had he been assigned to a BMI of 20 some time before baseline, his counterfactual outcome at the end of the study would have been necessarily equal to his observed outcome because there are many possible methods to assign someone to a BMI of 20."

Hernán here invokes an informal notion of intervention to argue against the consistency of BMI.

References

Nancy Cartwright. Modularity: It Can - and Generally Does - Fail. *Institute of Philosophy*, 2001. URL `https://sas-space.sas.ac.uk/986/`.

Patrick Forré and Joris M. Mooij. Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders. *arXiv:1807.03024 [cs, stat]*, July 2018. URL `http://arxiv.org/abs/1807.03024`. arXiv: 1807.03024.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL `https://www.nature.com/articles/ijo200882`.

David K Lewis. Causation. *Journal of Philosophy*, 1986.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl. Physical and Metaphysical Counterfactuals: Evaluating Disjunctive Actions. *Journal of Causal Inference*, 5(2), August 2017. doi: 10.1515/jci-2017-0018. URL `https://www.degruyter.com/view/journals/jci/5/2/article-20170018.xml`. Publisher: De Gruyter Section: Journal of Causal Inference.

Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL `https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml`.

Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

James Woodward. Causation and Manipulability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL `https://plato.stanford.edu/archives/win2016/entries/causation-mani/`.

Appendix: