Causal explanation with unobserved variables

—

We consider a disjunctive structure where each node X has an associated node term X_u, so that the structural equation is:

$$E := (A\&A_u) \vee (B\&B_u)$$

While the values of A and B are observed, the values of A_u and B_u aren't. The challenge is to develop an extension of the CESM that delivers causal scores even in the presence of uncertainty about the values of A_u and B_u.

Actual causation

One thing to consider first is that sometimes a variable value will just not count as a cause of the outcome. For example if E happens and $B = 1$ but $B_u = 0$, then intuitively neither $B = 1$ nor $B_u = 0$ was a cause of E. The CESM does not automatically weed out these non-causes, so we need a first step of processing that assigns the score of 0 to variables values that do not count as causes.

While there are formal theories of this categorical notion of causation, here I think we can go with our intuitions to determine what are the non-causes. For example in the example above $B_u = 0$ is not a cause of $E = 1$ because there is no conceivable situation where $B_u = 0$ is necessary for $E = 1$. And $B = 1$ is not a cause of $E = 1$ because the fact that $B_u = 0$ means that B was not able to 'reach out' to influence the outcome.

Computing a posterior distribution over unobserved variables

Another crucial step is inference: When we observe the value of A and B, we can update our probability distribution over what A_u and B_u can be in the actual world. For convenience we will denote this posterior distribution in an abbreviated form as:

$$Pr(A_u, B_u | A, B, E) = Pr_\alpha(A_u, B_u)$$

where the α denotes the fact that we're talking about the variables' values in the actual world. We can compute this posterior distribution in the normal way by using Bayes' rule.

Computing a CESM score by marginalizing over possible states of the world.

The CESM is defined for situations where we already know the full state of the world. To apply it to the present case (where this assumption doesn't hold), we must make some choices as to how to handle the uncertainty over A_u and B_u.

One intuitive way to do it is to compute a CESM score for each possible state of the actual world compatible with what we know, and then compute a weighted average of these scores, where the weights are the probabilities of the states of the world. For example to compute the CESM score for $A = a$, denoted $K(A = a)$, we compute:

$$K(A = a) = \sum_{A_u, B_u} K(A = a | A = a, B = b, A_u, B_u) Pr_\alpha(A_u, B_u)$$

where a and b are the actual-world values of A and B, and $K(V = v | \mathbf{X} = \mathbf{x})$ is the CESM score we would compute for $V = v$ if we knew that the actual-world values of $\mathbf{X}$ were $\mathbf{x}$.

Computing the CESM score for the unobserved variables introduces one additional complication: We typically don't know whether the variable has value 1 or 0. One intuition is that people will tend to say '$A_u = 1$ caused the outcome' if i) it is in fact likely that $A_u = 1$ in the actual world, ii) $A_u = 1$ has a high CESM score. One way to implement this is to multiply the CESM score by the probability of the variable value. Let us denote $K'(A_u = 1)$ the CESM score we would give to $A_u = 1$ if we were sure that A_u was in fact 1, and denote $K(A_u = 1)$ the score that accounts for the uncertainty. Then we have:

$$K(A_u = 1) = K'(A_u = 1) Pr_\alpha(A_u = 1)$$
$$= \sum_{B_u} K(A_u = 1 | A = a, B = b, A_u = 1, B_u) Pr_\alpha(B_u | A_u = 1) Pr_\alpha(A_u = 1)$$
$$= \sum_{B_u} K(A_u = 1 | A = a, B = b, A_u = 1, B_u) Pr_\alpha(A_u = 1, B_u)$$