Causal Inference, is just Inference: A beautifully simple idea that not everyone accepts

David Rohde
Criteo AI Lab,
Paris, France
d.rohde@criteo.com

Abstract

It is often argued that causal inference is a step that follows probabilistic estimation in a two step procedure, with a separate statistical estimation and causal inference step and each step is governed by its own principles. We have argued to the contrary that Bayesian decision theory is perfectly adequate to do causal inference in a single step using nothing more than Bayesian conditioning. If true this formulation greatly simplifies causal inference. We outline this beautifully simple idea and discuss why some object to it.

1 Introduction

Causal inference is often viewed as its own domain requiring concepts beyond standard probability and Bayesian decision theory. We think this complicated view is unnecessary. Bayesian decision theory automatically covers causal inference as a special case. Causal inference is complicated, not because new principles are needed but because probabilistic modelling in causal settings is difficult. Here we will show how simple Bayesian conditioning is sufficient to do causal inference and discuss why not everyone accepts the argument.

2 Bayesian Inference on Exchangeable observations

Imagine we measure an outcome on unit i, with binary outcome Y_i that received a binary treatment T_i. Furthermore, assume we have access to a dataset consisting of N different units i.e. our dataset is $Y_{1:N}$ and $T_{1:N}$. Furthermore we would like to set some future treatment T^* on another unit in the future. Our goal is to set T^* so that it will influence the outcome of Y^* and by convention we consider the outcome $Y^* = 1$ to be preferable to $Y^* = 0$. In other words the goal of our decision making problem is to determine how the treatment T^* influences the outcome Y^* and to set the treatment to maximize the probability that $Y^* = 1$.

We argue that the completely general algorithm to compute this probability is rather simple. To determine if we wish to treat $T^* = 1$ or not treat $T^* = 0$ we must specify a probabilistic model $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$, we then condition $P(Y^*|Y_{1:N}, T_{1:N}, T^*)$. Finally we compute: $\text{best } t = \text{argmax}_{t^*} P(Y^* = 1|Y_{1:N}, T_{1:N}, T^* = t^*)$. Notably, this algorithm is a straightforward application of Bayesian Decision theory, with the introduction of no novel notations or concepts to accommodate the causal aspect. Causal inference is often viewed as complex and difficult, "causation is not correlation" is a cliche of statistics. So our claim that causal inference can be reduced to computing a (Bayesian) conditional probability may be viewed with suspicion.

The point of view we develop here argues that causal inference is indeed difficult, but not because Bayesian conditioning is insufficient but rather because the task of probabilistically modelling $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ is difficult.

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

This modelling task is also difficult in ways that somebody familiar with using Bayesian modelling for associations might overlook. Let's consider some typical modelling assumptions that we might apply only to the observational part of the model (which is a more familiar problem to many) i.e. $P(Y_{1:N}, T_{1:N}) = P(Y_{1:N}, T_{1:N}|T^*) = \int P(Y_{1:N}, T_{1:N}, Y^*|T^*)dY^*$. Usually we will assume exchangeability (or conditional independence). This is done by introducing parameters, a general way to do this is:

$$P(Y_{1:N}, T_{1:N}) = \int P(\beta, \phi) \prod_n P(Y_n|T_n, \beta, \phi)P(T_n|\beta, \phi)d\beta d\phi \tag{1}$$

Using this model we can "fill in" missing parts of the observational data. e.g. if Y_N was missing then we could compute $P(Y_N|Y_1, .., Y_{N-1}, T_1, ..., T_N)$ but equally if T_N were missing we could compute $P(T_N|Y_1, .., Y_N, T_1, ..., T_{N-1})$. The conditional probability can be viewed as "causing you to think" - or as de Finetti puts it:

> I do not look for why THE FACT that I forsee will come about, but why I DO forsee that the fact will come about. It is no longer the facts that need causes; it is our thought that finds it convenient to imagine causal relations to explain, connect and forsee the facts. Only thus can science legitimate itself in the face of the obvious objection that our spirit can only think its thoughts, can only conceive its conceptions, can only reason its reasoning and cannot encompass anything outside itself. de Finetti (1975) [7]

The cause to think interpretation allows resolution of certain associations. For example observing Christmas cards might cause you to think it is Christmas even if they do not "cause" Christmas.

There are also more restrictive assumptions, one is the following construction based on the "regression assumption":

$$P(Y_{1:N}, T_{1:N}) = \int P(\beta)P(\phi) \prod_n P(Y_n|T_n, \beta)P(T_n|\phi)d\beta d\phi, \tag{2}$$

which introduces a further partial exchangeability assumption. According to Equation 1 pairs of Y, T may be permuted i.e. The probability remains the same if $Y_i = y_i, T_i = t_i, Y_j = y_j, T_j = t_j$ or if $Y_i = y_j, T_i = t_j, Y_j = y_i, T_j = t_i$ and all other elements are the same. Assuming exchangeability allows not only exchanging pairs but arbitrary numbers of permutations.

A further exchangeability constraint is implied by Equation 2 i.e if $T_i = T_j$ then you may permute Y_i and Y_j. One way to understand this assumption is that it is only possible to learn about the association between Y_i and T_i is by observing pairs of Y and T - semi-supervised learning based on only observing T_j without Y_j is not possible.

If we were to marginalize the model to contain only $T_{1:N}$ we have $P(T_{1:N}) = \int P(\phi) \prod_n P(T_n|\phi)d\phi$. Which assumes the elements of T_i and T_j are exchangeable.

A further important remark is that this assumption does not constrain any marginal $P(Y_i, T_i)$ but does constrain the joint over $P(Y_{1:N}, T_{1:N})$. This will become important when we address critiques of probability theory as able to solve causal inference problems.

Another possibility is:

$$P(Y_{1:N}, T_{1:N}) = \int P(\alpha)P(\lambda) \prod_n P(T_n|Y_n, \lambda)P(Y_n|\alpha)d\alpha d\lambda. \tag{3}$$

Similar to above this implies partial exchangeability i.e. if $Y_i = Y_j$ then you can permute T_i and T_j and It also implies exchangeability on the marginal $P(Y_1, .., Y_N)$

We can consider three different scenarios over $P(Y_{1:N}, T_{1:N})$:

1. A model that only assumes exchangeability over pairs of Y and T using the $P(Y_n|T_n, \phi, \beta)P(T_n|\phi, \beta)$ construction
2. A model that in addition to 1. assumes partial exchangeability of Y if T is the same using the $P(Y_n|T_n, \beta)P(T_n|\phi)$ construction
3. A model that reverses the assumptions in 2. i.e. assumes partial exchangeability of T if Y is the same using the $P(T_n|Y_n, \lambda)P(Y_n|\alpha)$ construction

It is worth noting these are different probabilistic models even if as $N \to \infty$ they all converge to the same $P(Y_{N+1}, T_{N+1}|Y_1, ..., Y_N, T_{1:N})$, the difference can be seen for example in considering if semi-supervised learning is possible. In the case of 2. Having access to measurements of T_j without the corresponding Y_j provides no information how Y_k is related to T_k and so semi-supervised learning is impossible [11] in the more general case of 1. semi-supervised learning may indeed be possible.

3 Causal Inference as Bayesian Inference

At this stage we move from predicting missing elements of $Y_{1:N}, T_{1:N}$ and return to the original causal problem of determining the treatment T^* in order to induce a preferred outcome on Y^*. This requires us to model: $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$. We need to connect the new outcome Y^* to the (to be chosen by us) treatment T^* and the observed data $Y_{1:N}, T_{1:N}$. If we base our model on Equation 1 we might arrive at:

$$P(Y_{1:N}, T_{1:N}, Y^*|T^*) = \int P(\beta, \phi)P(Y^*|T^*, \beta, \phi) \prod_n P(Y_n|T_n, \beta, \phi)P(T_n|\beta, \phi)d\beta d\phi, \quad (4)$$

which unfortunately is too general for any firm conclusion to be drawn and the details of the parameteric forms and priors matter even as $N \to \infty$. In contrast this extension of Equation 2 makes strong partial exchangeability assumptions and as a consequence allows (intersubjective) causal inference:

$$P(Y_{1:N}, T_{1:N}, Y^*|T^*) = \int P(\beta)P(\phi)P(Y^*|T^*, \beta) \prod_n P(Y_n|T_n, \beta)P(T_n|\phi)d\beta d\phi, \quad (5)$$

Intersubjectivity refers to the fact that Bayesian models that agree on exchangeability but otherwise differ can rapidly reach consensus. This is a consequence of the Bayesian law of large numbers i.e. if two Bayesians agree on exchangeability but otherwise have different priors then both will have a predictive distribution that rapidly converges to the observed frequency as $N \to \infty$.

If we adopt the assumptions in Equation 5 we then assume that if we set $T^* = t$, then Y^* is exchangeable with any Y_j if $T_j = t$. In practice this means by the Bayesian law of large numbers,that as $N \to \infty$; $P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*) \to$ empirical average of the subset of Y_j where $T_j = t$. This is the type of assumption we usually want to make when doing causal inference and this assumption is employed and appropriate after a well executed randomized control trial.

The partial exchangeability in scenario 3. where we use the $P(T_n|Y_n, \lambda)P(Y_n|\alpha)$ representation reverses the exchangeability and results in as $N \to \infty$; $P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*) \to$ empirical average of the of all Y. This is the situation where T does not cause Y, which is trivial - but usefully demonstrates the impact of different partial exchangeability relationships.

Unfortunately the assumption in Equation 5 often cannot be applied (or there is disagreement about if it can be applied) and only Equation 4 might be applied which implies no use-able partial exchangeability relationship. While Equation 4 is sufficient to make causal inference very dependent on assumptions - an alternative way to demonstrate the breakdown of any useful exchangeability result is to introduce a covariate into the model and then to discuss the impact of this covariate being hidden (an unobserved confounder). Making X the covariate the model becomes: $P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*)$ If we have:

$$P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*) = \int P(\gamma)P(\eta)P(\zeta)P(Y^*|X^*, T^*, \gamma)P(X^*|\zeta) \times \prod_n P(Y_n|X_n, T_n, \gamma)P(T_n|X_n, \eta)P(X_n|\zeta)d\gamma d\eta d\zeta, \quad (6)$$

but we only observe $Y_{1:N}, T_{1:N}$ - there is no exchangeability result that can be exploited and an intersubjective treatment effect cannot be learned - it is also reasonable to expect most individual Bayesians observing $Y_{1:N}, X_{1:N}T_{1:N}$ will not learn much about $P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*)$. Introducing an unobserved variable is just one way to show how exchangeability can break down. In statistical inference unobserved parameters are introduced to produce exchangeable probability models and are occasionally referred to as an indulgence in the strict "operational subjective" theory [10]. In causality unobserved confounders are introduced with the opposite purpose to destroy exchangeability and partial exchangeability between the the observed and future outcomes, but the introduction of a latent variable could equally be viewed as an indulgence.

When the covariate X is observed there are two plausible causally relevant ways a future Y^*, X^* may partially exchange with $Y_{1:N}$,$X_{1:N}$. Which results in Simpson's paradox [17]. The first of these is shown in Equation 6 with X observed, the second is given by:

$$P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*) = \int P(\gamma)P(\varpi)P(\varrho)P(Y^*|X^*, T^*, \gamma)P(X^*|T^*, \varpi) \times \prod_n P(Y_n|X_n, T_n, \gamma)P(X_n|T_n, \varpi)P(T_n|\varrho)d\gamma d\varpi d\varrho. \quad (7)$$

In the case of Equation 6 a partial exchangeability relationship exists between Y_j and Y^* so long as $X_j = X^*$ and $T_j = T^*$. In the case of Equation 7 a different partial exchangeability relationship exists between Y_j and Y^* and X_j and X^* so long as $T_j = T^*$.

4 Conclusion

Bayesian theory uses reasonable axioms of rational behaviour to show how we can use the knowledge of observed outcomes to update beliefs about other outcomes. It does not matter in principle if these observations are free form events, repetitions of a phenomena (allowing exchangeability) or are the outcome caused by a hypothetical intervention. To argue against this would require a critique of the axiom systems (See Appendix A).

It is however the case that once exchangeability is assumed as is possible in most purely observational studies the subtleties around exchangeable and partial exchangeable relationships between records can be mostly overlooked. When we must consider the causal outcome of an intervention this subtlety cannot be avoided and the probabilistic specification may be quite subjective. In this case different researchers will make a different causal inference, which is indeed a common situation when a high quality randomized control trial is not available.

It is also the case that a separate conditional probability must be computed[1] i.e. $P(Y^*, Y_{1:N}, T_{1:N}|T^* = 0)$ and $P(Y^*, Y_{1:N}, T_{1:N}|T^* = 1)$. Probability theory is entirely satisfactory to a) make causal assumptions and b) do causal inference via conditioning.

Alternative approaches separate statistical and causal inferences into separate steps. These steps involve estimation of a joint $\hat{P}(Y, T)$ and construct a causal effect as a transform of $\hat{P}(Y, T)$.

As mentioned not everybody accepts this methodology that uses probability (and partial exchangeability) both to encode associations and causal assumptions and uses only probabilistic conditioning to do the causal inference. Instead a two step procedure is adopted involving a statistical estimation of a (frequentist) distribution e.g. $\hat{P}(Y, T)$ and a causal step that explains if it is possible to recover the causal effect from $\hat{P}(Y, T)$. We argue that reducing the Bayesian $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ to the frequentist $\hat{P}(Y, T)$ obscures the partial exchangeability probability relationships that are fundamental

[1] We have no appetite to argue with anyone who sees this as an extension of probability theory.

to causal inference and requires the introduction of non-probabilistic methods both to encode causal assumptions and do causal reasoning in lieu of the simplicity and generality of probability theory. Not everyone agrees with us, and while some researchers are enthusiastic about or formulation it is rejected by key thinkers in the causal community.

In Appendix B we discuss non-probabilistic approaches to causality that separate inference into causal and statistical steps. Appendix C responds to some of the criticism and provides key references.