



PROJECT MUSE®

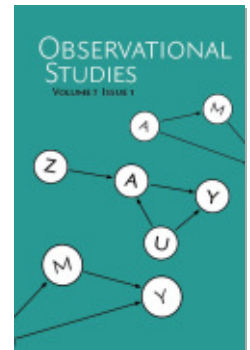
Causal Modelling: The Two Cultures

Elizabeth L. Ogburn, Ilya Shpitser

Observational Studies, Volume 7, Issue 1, 2021, pp. 179-183 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0006>



➔ *For additional information about this article*

<https://muse.jhu.edu/article/799731>

Causal Modelling: The Two Cultures

Elizabeth L. Ogburn

eogburn@jhsph.edu

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

Baltimore, MD 21205 USA

Ilya Shpitser

ilyas@cs.jhu.edu

Department of Computer Science

Johns Hopkins University

Baltimore, MD 21218 USA

Abstract

We offer descriptive and normative standards for the principled pursuit of causal inference. These standards address critiques of both the algorithmic and the data modeling cultures identified in (Breiman, 2001), and provide a fruitful synthesis of both cultures. We contrast the resulting "cautious causal inference" with overly optimistic methods inspired by algorithmic data analysis methods prevalent in machine learning, as well as older approaches to causal modeling that employ overly restrictive parametric models.

Keywords: Causal Inference, Nonparametric Identification, Machine Learning

Breiman's influential 2001 paper posited a dichotomy in statistics between what he called the *data modelling* culture and the *algorithmic modelling* (i.e. machine learning) culture. Breiman described the data modelling culture as being concerned with understanding mechanisms by which the data is generated, while the algorithmic culture treats the underlying data generating mechanism as unknown and perhaps irrelevant. As a number of original replies to Breiman's paper noted, Breiman's description of the data modelling culture includes causal modelling, although his original description did not use explicitly causal language.

Breiman's critique of the data modelling culture can be boiled down to two observations. First, data modelling, as Breiman saw it at the time, primarily used parametric models. Breiman viewed these models as unsuitable for the sorts of applications he had in mind, due to issues arising from misspecification and poor model fit. Second, Breiman correctly noted that the emphasis on understanding true mechanisms underlying data generation is not actually needed (and may be counterproductive) in a wide variety of problems, such as those pertaining to prediction, clustering, or dimension reduction.

While Breiman himself clearly favored the algorithmic modelling culture, it is itself open to critique, with some of it already articulated in the original 2001 replies to Breiman. In particular, understanding causal mechanisms is crucial in many areas of empirical science and in rational decision making, as evidenced by the ubiquity of randomized controlled trials and the increasing adoption of causal inference methods for observational data. The

enterprise of causal inference is sufficiently subtle that algorithmic modelling ideas alone, agnostic as they are to true underlying mechanisms, are insufficient for obtaining valid answers. In causal inference, this agnosticism can lead to what we call *sanguine causal modelling*: the use of mathematically convenient identifying assumptions, combined with optimism that these convenient assumptions will happen to hold in practice. These kinds of assumptions can be unnecessarily strong, and difficult to reason about.

Thus, we disagree with Breiman on the primacy of the algorithmic modelling view of data analysis, especially for causal inference. However, many of Breiman’s critiques of the data modelling culture remain valid, particularly for data science communities that work exclusively with strong, often parametric models that are likely to be misspecified in practice.

The culture of *cautious causal modelling*, which shuns opaque assumptions on the fear that they could result in incorrect conclusions, has been sensitive to both types of critiques. Cautious causal modellers identify causal effects with assumptions that are transparent and only as strong as necessary. To estimate causal effects, cautious causal modellers employ flexible statistical models and robust methods in order to avoid falling prey to model misspecification. To put it another way, cautious causal modelling aims to synthesize the best aspects of the data modelling and algorithmic modelling cultures by clearly specifying the parameter of interest and the modelling assumptions under which it will be identified, as data modelers would, while also using, to the extent possible, the flexible estimation methods developed by communities adhering to the algorithmic modelling view of data analysis.

We will summarize the workflow of cautious causal inference and how this workflow achieves this synthesis. But first, some terminology. *Potential* or *counterfactual* outcomes are the outcomes that we would have observed, possibly contrary to fact, had we been able to intervene to set treatment or exposure to a particular user-specified value. They are often denoted $Y(x)$ for a treatment or exposure X set to value x . Potential outcomes are only partially observed: for binary X we observe $Y(1)$ but not $Y(0)$ for units with $X = 1$, and we observe $Y(0)$ but not $Y(1)$ for units with $X = 0$. The *full data* is the random vector $(Y(1), Y(0), X, C)$ for covariates C , while the observed data (Y, X, C) is a coarsening of the full data with $Y = X * Y(1) + (1 - X) * Y(0)$. A *causal estimand* can be any functional of the full data, e.g. a contrast between $E[Y(1)]$ and $E[Y(0)]$. Absent assumptions, we cannot learn about causal estimands from the observed data, because they are functionals of (partially) unobserved potential outcomes. We say that a functional of the full data is *identified* by the observed data, under a set of assumptions, if the functional reduces to a functional of the observed data alone. This functional of the observed data is the *identifying functional*, and it is the *statistical estimand* that we will estimate and perform inference about in order to learn about the causal estimand that it identifies.

Now we can describe the crucial features of cautious causal modelling. Cautious approaches to causal inference:

1. define the causal estimand(s) of interest in terms of the full data and without reference to any assumptions or models, which will be invoked later;
2. propose a set of assumptions that can be justified by existing research, domain expertise, or that are mathematically unrestrictive;

3. communicate the assumptions transparently, e.g. offering intuition for any technical assumptions and explaining how a practitioner could reason about when they hold or are violated;
4. deduce whether and how the observed data identifies, or partially identifies, the causal estimand given the assumptions;
5. estimate (or otherwise use data to learn about, e.g. by deriving bounds on) the observed data estimand using only justifiable or unrestrictive models;
6. assess the robustness of the conclusions from (5) to violations of causal and modelling assumptions.

Step (1) separates the scientific question from the quantitative methods that will be used to answer it, ensuring that we do not fall into the trap Breiman himself succumbed to in his focus on prediction problems, namely retrofitting the question to match preferred methods. Together steps (1) and (2) ensure that researchers do not reflexively fit a popular parametric model and interpret a coefficient as a causal effect (as Breiman’s data modeller might do). One of Breiman’s primary concerns about the data modelling culture was over-reliance on parametric models which, if they are “a poor emulation of nature,” may lead to incorrect conclusions. For this very reason cautious causal modellers are skeptical of parametric (or overly strong) identifying assumptions in step (2), unless they can be justified as Breiman describes in his response to Cox: “if the science of the mechanism producing the [ed: full] data is well enough known to determine the model apart from estimating parameters.” In fact, the use of parametric models in step (2) is even more troubling than their use for estimation, which is what concerned Breiman. Typically *no* goodness-of-fit tests, not even the problematic ones Breiman discussed, are readily available for the fit of a parametric model or assumption to the full, rather than the observed, data. This is because it is difficult to discern what restrictions on the observed data are implied by restrictions on the full data.

Steps (3) and (4) are what separate causal from other kinds of statistical inference. Since causal inferences are complex counterfactual claims about how outcomes would have been different under hypothetical changes to treatment, there is no reason to believe these claims unless researchers are able to reason clearly about whether and when the assumptions on which they rest might be expected hold. A cautious approach will prioritize assumptions that are less restrictive, so that they are more likely to hold in practice, and assumptions that are substantive rather than technical, to facilitate step (3). What makes an approach ultimately useful is that the assumptions are transparent and amenable to being reasoned about.

A hallmark of sanguine causal modelling is putting step (4) before step (2), i.e. assuming that an estimand is identified, and then back-engineering assumptions to justify identification. This can render the methods useless in practice, if the assumptions are unlikely to hold or impossible to reason about. Sometimes it results in identifying assumptions that essentially assume the consequent, i.e. that assume that *a particular functional of the full data is equal to a particular functional of the observed data*. Particularly concerning is when methods that rely on restrictive, technical assumptions are marketed as being widely applicable in practice. When this kind of method is applied to real data, researchers have

no way of knowing whether or not the assumptions hold and causal inferences are actually licensed.

Step (5) has evolved considerably since Breiman’s paper was published. The use of parametric models for estimation in step (5) has been a growing concern among cautious practitioners of causal inference; over the past decade much of the research in causal inference has focused on how to use nonparametric—even blackbox—methods in the service of Step (5) (Chernozhukov et al., 2017; Hill, 2011; Kennedy, 2020; Liu et al., 2020; Zheng and van der Laan, 2011). The causal community has come to embrace semi- and non-parametric estimation methods that allow flexible nuisance models to be used while still obtaining estimators with desirable parametric properties (such as asymptotic normality and \sqrt{n} -consistency) for the target parameter. In addition, semi-parametric estimators for many popular target parameters in causal inference have product biases, resulting in multiple robustness properties, where the estimator remains consistent even if some of the nuisance models are arbitrarily misspecified.

Ideally, researchers seeking to draw causal conclusions will triangulate different types of analyses with independent weaknesses (Rosenbaum et al., 2017). When that is not possible, step (6) involves formal and informal sensitivity analyses: using many different parametric models to assess sensitivity to instability and model fit; assessing to what extent the findings are robust to violations of the causal identifying assumptions. For the specific assumption of *no unmeasured confounding*, there is a large literature on formal sensitivity analysis techniques (Dorie et al., 2016; Franks et al., 2019; Liu et al., 2013; McCandless et al., 2007; Rosenbaum, 2014). This directly addresses Breiman’s concern that data modelling approaches are not robust to model misspecification or to instability.

The synthesis of the data modeling and algorithmic modeling cultures, due to the cross-pollination of ideas from machine learning and causal inference, is only just beginning. The number of causal contributions to machine learning conferences seems to grow exponentially from conference to conference, and we are witnessing increasing adoption of machine learning estimation methods in causal inference. We look forward to taking stock, in another 20 years, of how far both fields have come from Breiman’s 2001 characterization.

References

- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.
- Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.
- AlexanderM Franks, Alexander D’Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.
- Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

- Edward H Kennedy. Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics*, 16(1), 2020.
- Lin Liu, Rajarshi Mukherjee, James M Robins, et al. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):518–539, 2020.
- Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14(6):570–580, 2013.
- Lawrence C McCandless, Paul Gustafson, and Adrian Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347, 2007.
- Paul R Rosenbaum. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.
- Paul R Rosenbaum et al. The general structure of evidence factors in observational studies. *Statistical science*, 32(4):514–530, 2017.
- Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.