Statistical Science
2020, Vol. 35, No. 3, 430–433
https://doi.org/10.1214/20-STS772
Main article: https://doi.org/10.1214/20-STS721

Comment: Invariance and Causal Inference

Stefan Wager

The problem of distinguishing causal effects from noncausal correlations is one of the oldest and most challenging questions in statistics. In recent years, Professor Bühlmann and co-authors have outlined new methodology for estimating causal effects that starts from an invariance postulate: A set of variables X is causally relevant to an outcome Y if the distribution of Y conditionally on X, $\mathcal{L}(Y \mid X)$, is invariant across all relevant environments. This hypothesis then leads to statistical methodologies that seek causal effects by fitting models that are robust across numerous environments (Peters, Bühlmann and Meinshausen, 2016, Rothenhäusler et al., 2019). The present paper, generously prepared by Professor Bühlmann, is an enlightening summary of this groundbreaking line of work and a valuable addition to the literature.

This invariance hypothesis presents a marked and thought-provoking departure from the currently dominant paradigm for understanding causal effects in epidemiology and econometrics, which defines causal effects in terms of potential outcomes and emphasizes the role of experimental design in identifying causal effects (Neyman, 1923, Holland, 1986, Robins and Richardson, 2010, Rubin, 1974, Rubin, 2005). In general, the potential outcomes based approach allows treatment effects to vary arbitrarily with both observed and unobserved features and is focused on defining, identifying and estimating various (weighted) treatment effect functionals under minimal assumptions. Characterizing how the invariance hypothesis fits into the potential outcomes framework is important to understanding how the results of Peters, Bühlmann and Meinshausen (2016) and Rothenhäusler et al. (2019) connect to more classical approaches.

Potential outcomes and weighted treatment effects. The earliest application of the potential outcomes framework was Neyman's analysis of the randomized controlled trial. In this setting, we are interested in measuring the effect of a binary treatment W_i on a real-valued outcome Y_i. We posit the existence potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome the ith observation would have experienced had they received treatment assignment 0 or 1, respectively, such that $Y_i = Y_i(W_i)$, and then define the sample average treatment effect[1]

$$(1) \qquad \tau_{\mathrm{SATE}} = \frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i(1) - Y_i(0)\bigr).$$

The seminal result of Neyman (1923) is that, if the treatment assignment W_i is randomized, that is, the treatment assignment is exchangeable and $\{W_i\}_{i=1}^n \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}_{i=1}^n$, then we can construct an unbiased estimate of τ_{SATE} without assumptions: No modeling assumptions are made on the potential outcomes $Y_i(w)$, and in fact the potential outcomes may even be taken as deterministic such that only W_i is random.[2] In particular, it is not necessary to assume that the causal effect is the same for each unit, for example, that $Y_i(1) - Y_i(0) = \tau$ for some shared (or invariant) causal parameter τ.

Starting with Rubin (1974), there has been considerable interest in generalizing the ideas of Neyman (1923) beyond the randomized controlled trial, and in developing appropriate treatment effect estimators that remain justified without making structural assumptions on the per-unit treatment effects $Y_i(1) - Y_i(0)$. One setting that has received considerable attention is that of Rosenbaum and Rubin (1983), where treatment assignment W_i is not randomized, but we observe covariates X_i such that W_i is as good as random after we condition on them, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$. Under an IID sampling model, the semiparametric efficient variance V for estimating the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ can be written in terms of the propensity score $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$ (Hahn, 1998, Robins and Rotnitzky, 1995),

$$V = \mathrm{Var}\bigl[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]\bigr] + \mathbb{E}\biggl[\frac{\mathrm{Var}[Y_i(0) \mid X_i]}{1 - e(X_i)} + \frac{\mathrm{Var}[Y_i(1) \mid X_i]}{e(X_i)}\biggr],$$

and efficient estimators satisfy $\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, V)$.

One complaint about this result, however, is that V scales with the inverse of the propensity score, and can get quite large if we have poor overlap (i.e., $e(X_i)$ can get

Stefan Wager is Assistant Professor of Operations, Information and Technology, and Assistant Professor of Statistics (by courtesy), Graduate School of Business, Stanford University, Stanford, California 94305, USA (e-mail: swager@stanford.edu).

[1] One major assumption here is that of no interference, that is, that W_i only affects the outcome of the ith unit (Imbens and Rubin, 2015). For a discussion of potential outcomes modeling under interference; see Basse, Feller and Toulis (2019), Hudgens and Halloran (2008) and references therein.

[2] These results can be considerably generalized. For example, Ding, Feller and Miratrix (2019) and Lin (2013) for a discussion of regression adjustments in this setting.