Statistical Science
2020, Vol. 35, No. 3, 427–429
https://doi.org/10.1214/20-STS768
Main article: https://doi.org/10.1214/19-STS721

Comment: Invariance, Causality and Robustness

Vanessa Didelez

I would like to congratulate Peter Bühlmann on the honor of being invited to give the Neyman Lecture. Jointly with a number of co-authors of recent papers, he has produced a substantial and thought-provoking body of work in recent years around the concept of invariance. His achievement is two-fold: He extends causal reasoning to involve prediction under new environments; after several decades of existing research in the field of causal inference; see early work in the 1970s and 1980s by Rubin, Robins, Pearl, Spirtes and colleagues and the ensuing explosion of work on this topic in bio-medical statistics, epidemiology, computer science, sociology and political science—this is a *novel angle* on causal inference, using data in a different way with an original target of inference so far undervalued in the causal inference literature. Vice versa, he demonstrates how causal reasoning is important to predictive modelling. It is a particular achievement of Bühlmann to have brought key ideas and concepts of causality and causal inference to the attention of mainstream statistics. This is not least due to linking causal ideas, such as invariance (also known as *stability* (Dawid and Didelez, 2010)), with fundamental concepts of traditional statistical inference, such as worst-case risk optimization.

In the following, I will review the differences and similarities of 'classical' causal inference and Bühlmann's approach.

CAUSAL INFERENCE, BIAS AMPLIFICATION AND PREDICTION

I would like to discuss some of the ideas in Bühlmann's paper by attempting to relate them to a phenomenon known in the bio-medical/causal inference literature as 'bias amplification' (Pearl, 2010, Middleton et al., 2015, Ding, VanderWeele and Robins, 2017). Consider a simple linear SEM where

$$Y = \beta X + \alpha H + \epsilon,$$

and where A is a valid instrumental variable for the effect β of X on Y (as in Bühlmann's Figure 6 with no $A \to H$ and no $A \to Y$ edges, see Figure 1(a) in this commentary). The classical aim of causal inference is to to estimate β: we may be interested in β because under the above SEM this parameter represents the effect on Y of fixing X at x versus fixing it at $x+1$, that is, the average causal effect $\beta = \mathbb{E}(Y|\text{do}(X = x+1)) - \mathbb{E}(Y|\text{do}(X = x))$ (due to linearity and no interaction, the marginal and the conditional average causal effects are the same in this special case; but we must not forget that this does not hold for more general models[1]).

In the above model, we know that (i) a linear regression of Y on X results in a biased estimator for β due to the hidden confounder H, unless the $H \to X$ or $H \to Y$ relations vanish; (ii) using A as an instrument to perform two-stage least squares (2SLS) yields a consistent estimator of β; (iii) regressing Y on both X and A (or partialling out A first) typically results in *even more* bias than approach (i). This last phenomenon is known as 'bias amplification' (Pearl, 2010). Intuitively, the amplification occurs because including the IV A as additional regressor explains away some of the 'free' variability in X, with the variability due to H remaining, and hence amplifying the bias due to unobserved confounding by H (Greenland and Pearl, 2011).

When using anchor regression, (i) corresponds to $\gamma = 1$, (ii) to $\gamma = \infty$, and (iii) to $\gamma = 0$. Hence, when A is a valid instrument, we can roughly say the larger γ the less bias we have in estimating β; at the same time (due to Theorem 4.1) for large γ we minimise worst-case prediction risk under large shift perturbations but not under small shift perturbations.

Consider now the case where A is not a valid instrument (see Figure 1(b) in this commentary) and

$$Y = \beta X + \alpha H + \xi A + \epsilon,$$

with $\xi \neq 0$ (note that under the shift perturbations ξA is replaced by the shift v). In this case, we know (i′) a linear regression of Y on X results in a 'doubly biased' estimator for β due to the hidden confounder H and the hidden (unused) confounder A; (ii′) using A as an instrument to perform two-stage least-squares will also yield a biased estimator of β as A is now not a valid IV anymore; (iii′)

Vanessa Didelez is Professor, Leibniz Institute for Prevention Research and Epidemiology—BIPS, and Faculty of Mathematics and Computer Science, University of Bremen, Germany (e-mail: didelez@leibniz-bips.de).

[1] Much of the causal inference literature is concerned with robust estimation of a marginal causal effects under *considerably weaker* parametric assumptions than a linear SEM.

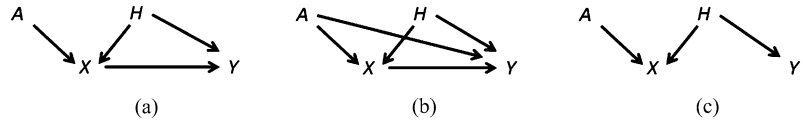

FIG. 1. *Causal DAGs, where (a) A is a valid IV for the effect of X on Y in the presence of unobserved confounding by H; (b) A is itself a confounder of the effect of X on Y and not a valid IV anymore; (c) A and X (with H still hidden) are useful for predicting Y in an unchanged environment but have both no causal effect on Y.*

finally, regressing Y on both X and A should result in a less biased estimator of β than (i′). The amount of bias and the extent to which the IV is invalid depend on ξ and the strength of the $A \to X$ relation; when ξ is close to zero, then (ii′) is less biased, when it is further away from zero then (iii′) will be better for estimating β. What happens to amplification bias which occurred for $\xi = 0$? For $\xi \neq 0$, approach (iii′) will still amplify the bias due to confounding by H whenever the inclusion of A as regressor reduces the unexplained variability in X more than the unexplained variability in Y, that is, this also depends on the $A \to X$ strength. So, interestingly, depending on the $A \to Y$ strength and the $A \to X$ strength, there must be a tipping point where using A as IV versus including it as a regressor becomes less beneficial with view to bias in estimating β.

How does this relate to anchor regression? Above, I was considering estimation of the causal effect parameter β. Anchor regression instead focusses on minimising worst-case prediction risk across different and new environments. As I understand it, estimation of the causal parameter β is most relevant for predicting Y in new environments where X is made *completely independent* of H, which corresponds to the truncation principle (Spirtes, Glymour and Scheines, 2000)—these are not the shift perturbations of Section 4.3 for which anchor regression has the optimality property of Bühlmann's Theorem 4.1. While causal reasoning about the relation between X and Y typically assumes that we can fix the value of X itself by an intervention, the shift perturbations instead formalise that the value of A is changed or replaced by the new environment and that this then shifts all other variables. Note that under a shift perturbation X retains an association with H so that the biased estimator from (i′) may still be useful for prediction in such a new environment. I will discuss the prediction aspect next.

PREDICTION AND CHANGING ENVIRONMENTS

A key insight underlying invariant prediction and anchor regression is to make explicit the need to be clear whether a prediction task is aimed at the *same* or at a *different* environment; obviously, the quality of prediction in a new environment depends on what is known, or can be inferred, about how the new environment is different from the training environment(s). Any background knowledge on stable aspects across new and old environments should be exploited. The invariance principle is much more general than causal relations; but if we are prepared to assume an underlying causal structure, then we know which invariances to expect under what type of perturbations or interventions.

Consider the standard task of predicting Y under *unchanged* environments. To illustrate this, assume that A is a valid IV and that X has in fact no causal effect (Bühlmann's Figure 6 with no $X \to Y$, no $A \to H$ and no $A \to Y$ edges, see Figure 1(c) in this commentary). In the absence of measurements on H, to minimise prediction error we should use *both* X and A as regressors, that is, use method (iii). This is because X is informative for H, and hence useful for predicting Y; and A is also predictive of Y due to its conditional association with Y given the collider X, that is, together with X the variable/indicator A becomes informative for H, and hence for Y; consequently, in a regression of Y on (A, X) both variables will typically have nonzero regression coefficients despite none of them being a cause of Y in Figure 1(c). In terms of anchor regression, the case of new environments being the same as the old environments occurs when the shift perturbations are zero; hence, the optimal choice is $\gamma = 0$ which coincides with the above reasoning and shows that for ordinary prediction (under unchanged environments) any conditionally associated predictors are relevant. In summary, for pure prediction an IV would simply be used as a regressor instead of 2SLS; bias amplification is not an issue as the aim is not to consistently estimate a causal effect.

The predictive role of A and X holds while the data generating process for (X, H, Y) under different environments A remains the same, but they may lose this property under certain changes. For instance, as discussed earlier, in a different experimental environment we might change the way how X is generated and make it independent of H; then X becomes entirely uninformative for H, and hence for Y, and similarly A will no longer be informative for Y given X. If X is not a cause of Y (Figure 1(c)) it is useless for prediction under such modified environments. Shift perturbations lie somewhere in between: with no shift we have unchanged environments and with large shifts we get closer to an experimental setting (if the shift affects X). What cannot be represented by shifts is an intervention that fixes X at some value near

its mean. Shifts do not render the variables of an SEM entirely independent of their graph parents, but they weaken the dependence. While the classical task of prediction (in unchanged environments) appears almost the 'opposite' of causal inference and can in principle easily be automatized, the novel task of prediction under changing environments require some careful causal thinking—to what extent it can also be automatized is a difficult question, and I would call for caution. A key structural assumptions underlying anchor regression, to be justified in any given application, and hence not automatic, is that the anchor A must be exogenous, that is, not itself causally affected by other variables in the system. For example, if the different environments are different hospitals, and it depends on the socioeconomic background of patients which hospital they attend, then it is unlikely that 'hospital' can be used as anchor.

A DEFINITION OF CAUSALITY?

In Section 6, Bühlmann discusses whether causality could be *defined* in terms of *invariance*. I would like to put the emphasis differently, namely on the importance of formalising the *changes* relevant to the research question. Classically, causal inference wants to asses effects of *interventions*, that is, quite extreme changes to a system. For instance, we might ask how health would improve in a population if *everyone stopped* smoking as could be enforced by a complete ban of cigarettes. The importance of well-defined interventions for practically useful causal analyses in epidemiology in terms of targets for action has received much attention (Hernan, 2016). In contrast, it has sometimes been criticised that humans cannot in practice intervene in many causal relations and, therefore, a definition of causation in terms of effects of interventions is besides the point (Pearl, 2009). However, in my view it is important to envisage *some change*, like the new environments of Bühlmann, in order to clarify the need for, and the aim of, causal inference. This change does not need to be 'human-made', it could be due to different or changing circumstances. Bühlmann's approach nicely illustrates this: if we want to simply predict Y under unchanged environments ($\gamma = 0$), we can use any associations (see earlier example in the context of Figure 1(c) above); but if we want to make predictions for changing environments we better start to apply causal principles and look for exogeneities. A key role falls to defensible assumptions about when what is, and what is not, invariant or stable under those changes; a thorough and deep understanding of these concepts is essential.

REFERENCES

DAWID, A. P. and DIDELEZ, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* **4** 184–231. MR2740837 https://doi.org/10.1214/10-SS081

DING, P., VANDERWEELE, T. J. and ROBINS, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* **104** 291–302. MR3698254 https://doi.org/10.1093/biomet/asx009

GREENLAND, S. and PEARL, J. (2011). Adjustments and their consequences—collapsibility analysis using graphical models. *Int. Stat. Rev.* **79** 401–426.

HERNAN, M. J. (2016). Does water kill? A call for less casual causal inferences. *Ann. Epidemiol.* **26** 674–680.

MIDDLETON, J. A., SCOTT, M. A., DIAKOW, R. and HILL, J. L. (2015). Bias amplification and bias unmasking. *Polit. Anal.* **24** 307–323.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PEARL, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the* 26*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-10) 417–424. Morgan Kaufmann, San Mateo, CA.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675