Statistical Science
2020, Vol. 35, No. 3, 434–436
https://doi.org/10.1214/20-STS797
Main article: https://doi.org/10.1214/19-STS721

Rejoinder: Invariance, Causality and Robustness

Peter Bühlmann

Abstract. We sincerely thank Vanessa Didelez and Stefan Wager for their insightful and inspiring comments. Their views and thoughts on the topic of my article are of great value and truly contribute to put it into greater perspective.

Key words and phrases: Anchor regression, causal regularization, distributional robustness, heterogeneous treatment effects, instrumental variables regression, random forests.

Vanessa Didelez's comments largely focus on Anchor regression and causal regularization, except at the end where she expresses her nice philosophical thoughts on the definition of causality; Stefan Wager raises fundamental issues on heterogeneity, stability and external validity. According to the alphabetical order of the discussants, we first respond to Didelez and then to Wager.

1. CHOOSING THE AMOUNT OF CAUSAL REGULARIZATION, AMPLIFICATION BIAS AND SPECIFICATION OF ANCHORS

Didelez explains in a very clear way what Anchor regression (Rothenhäusler et al., 2018) does and what it doesn't. I am repeating: she starts by considering the problem whether one should adjust for anchors A (ordinary least squares including A, OLS+A) or whether one should project onto the linear span of A (two-stage least squares, 2SLS); the other standard option is ordinary least squares versus X only (OLS) and discarding the anchor variables, perhaps most prevalent if the anchor A encodes different sub-populations or environments but one treats the data as homogeneous. Anchor regression includes all of these options for different amount of causal regularization: OLS+A with $\gamma = 0$, 2SLS with $\gamma = \infty$ and OLS with $\gamma = 1$. Didelez nicely points out that there could be amplification bias with OLS+A ($\gamma = 0$) depending whether A is an instrument or a "not too severely invalid" instrument.

An additional and often chosen approach is as follows. The variables A would encode the first few principal components of X, and hence A would *not* be exogenous: then OLS+A ($\gamma = 0$) is adjusting for these first principal components and aims to remove hidden confounding bias in estimating the causal parameter β. This common practice in applied statistics (e.g., Novembre et al., 2008) can be justified under the assumption of *"dense confounding"* where the hidden variables H affect most of the X components (Ćevid, Bühlmann and Meinshausen, 2018); see also closely related work by, for example, Chandrasekaran, Parrilo and Willsky (2012), Shah et al. (2018), Guo, Ćevid and Bühlmann (2020). The theoretical and methodological arguments are different since A is now a proxy for the hidden latent confounder H, very different from a valid instrument and not exogenous. In addition, the data consists only of X and Y and there are no heterogeneities. However, from an operational view point, Anchor regression includes this procedure as well.

The nice exposition of Didelez and the example above about adjustment with the few first principal components illustrates that choosing the amount of causal regularization, the parameter γ, depends very much on the underlying scenarios, namely the nature of the anchors and the structure and quantitative relations among the underlying variables. We have no fully satisfactory proposal yet for choosing γ. As of today, we suggest the following:

(i) If the anchors are assumed to be exogenous, that is we would have additional external information: then γ is the strength of future perturbations we want to insure against, in terms of prediction (see Theorem 4.1 in the main article; essentially, γ is the factor of strength of future perturbations relative to the observed heterogeneity in the data).

(ii) If interested in prediction for new environments, one may rely on leave environment out cross-validation and optimize the *worst-case* prediction error on left-out environments. This strategy implicitly assumes that the

Peter Bühlmann is Professor of Statistics, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

new unseen environment is "within the range of environments" we have observed in the data.

Both of these proposals rely on predictive aspects.

Reducing Variance

By good reasons, Didelez mentions issues with bias. But even in well posed situations with valid instruments, the 2SLS estimator has a very large or even infinite variance. In such a situation, Anchor regression with relatively large values of γ will substantially reduce variance at the expense of a (typically slight) increase of bias for estimating the causal parameter β. Thus, even with valid instruments, it pays off to do causal regularization for improving the MSE of estimating the causal parameter β. In fact, Anchor regression exactly corresponds to the K-class estimators (Theil, 1958) which generalize and improve over 2SLS; in fact, the latter has been an invention by the same scientist Theil (1953) a few years earlier. Jakobsen and Peters (2020) provide new strategies to choose the amount of causal regularization γ from the perspective to optimize the MSE for estimating the causal parameter β.

2. DIFFERENT INTERVENTION MODELS

Didelez nicely points out that the type of intervention matters in terms of predictive aspects. Anchor regression considers shift interventions: they are very different from do-interventions. The former is perhaps better described as a perturbation: it does not change the structural parts of the system (the graph remains the same) but is shifting the involved random variables. The latter is an intervention which also changes the structure (the graph) according to the truncation principle: it is a model with a rather strong "decoupling" effect. The choice of the intervention *model* matters, not only mathematically: one should ask also here for what kind of application a certain model is a reasonable approximation.

In contrast to Anchor regression, Invariant Causal Prediction (Peters, Bühlmann and Meinshausen, 2016) works for a large class of intervention models, but makes more stringent assumptions that Y is not allowed to be directly intervened on. There is a large territory for further exploiting the characteristics and properties of different intervention models.

3. EXTERNAL VALIDITY AND TRANSFER LEARNING

Wager provides very insightful comments on the potential outcomes framework. Indeed, it has many advantages: for example, it does not require an i.i.d. sampling assumption in general and each unit i may exhibit a probabilistically different treatment effect $Y_i^{(1)} - Y_i^{(0)}$ on a response of interest. As Wager points out, the potential outcomes framework becomes cumbersome in the multivariate case with many (treatment) variables.

Wager nicely phrases that the potential outcomes framework strives for "well-defined weighted in-sample in-study average treatment effects under minimal assumptions". The invariance paradigm on the other hand naturally leads to transfer learning and external validity as discussed in the main article and also in other work (Rojas-Carulla et al., 2018, Pfister et al., 2019). Under some (strong) assumptions, the concepts of potential outcomes, causality in structural equation models or also the approach based on invariance or stability coincide (Dawid and Didelez, 2010, Richardson and Robins, 2013, Peters, Bühlmann and Meinshausen, 2016, Dawid, 2020).

Increased Replicability

Stability and invariance lead to better replicability and external validity. Eventually, a scientific study is only interesting if it generalizes to a new individual or unit which was not included in the study. It is important to improve the degree of replicability for new individuals arising from somewhat different (sub-) populations. Such increased replicability, due to invariance or stability, can be mathematically described and exploited on real data: the diluted causal parameter $\beta(\gamma \to \infty)$ (Section 4.3.2 in the main article) has such a replicability property when the new population arises from shift perturbations on the training population. The details are given in Rothenhäusler et al. (2018), Section 3.1, Section 5.1, Figure 11. Of course, the direct causal parameter in a structural equation model is also replicable if the perturbation generating the new environment does not affect the response Y directly.

3.1 Heterogeneous Treatment Effects and the California GAIN Study

Wager provides a very nice empirical illustration wit the California GAIN study. The average treatment effects (from the population outcomes framework) are not stable over the different environments, which are in this context the different counties. But when considering conditional treatment effects, they appear to be stable. The conditional treatment effect formulation can be brought into the perspective of the invariance framework as follows (and indeed, we haven't highlighted this so far).

Consider an anchor regression type model:

$$Y \leftarrow T\tau(X) + (1-T)\mu(X) + M_Y A + B_{Y,H} H + \varepsilon_Y,$$
$$X \leftarrow M_X A + B_{X,H} H + \varepsilon_X,$$
$$H \leftarrow M_H A + \varepsilon_H,$$

where $T \in \{0, 1\}$ is a binary treatment variable, $\tau(\cdot)$ is the direct conditional treatment effect function of the covariates, $\mu(\cdot)$ is the conditional mean function of the covariates for the control (nontreated) cases, and A is an

exogenous anchor, say encoding different environments. This is the analogous model to the nonlinear anchor regression model in equation (5.1) in the main article. We could then use nonlinear Anchor regression to estimate a "stabilized version" of $\tau(\cdot)$ and $\mu(\cdot)$, for example, with a boosting-type optimization scheme and using causal forests as weak learner (Wager and Athey, 2018). Such an anchor regression approach would not rely on the assumption that the treatment is randomized conditional on the covariates. Instead, it uses the anchors to stabilize and robustify against shift perturbations in order to improve external validity; in the special case where A is a valid instrument, a nonlinear IV technique leads to the true causal conditional treatment effect function $\tau(\cdot)$ (when conditioning on X), even in presence of potential unobserved confounding. We agree with Wager that much more should be done to bring the notions of heterogeneity of covariates and invariance together in a fruitful way: of course, as Didelez has pointed out, the issue of distinguishing anchors and covariates is very relevant here as well, perhaps even more so.

4. WHAT IS CAUSALITY?

Didelez' comments on "A Definition of Causality?" are very well-thoughtout and inspiring, also from a philosophical point of view! Highly recommended to read.

ACKNOWLEDGMENTS

We thank the Guest Editors and the Editor for the opportunity of presenting our work as a discussion paper.

This work has been funded in part by the European Research Council under the Grant Agreement No. 786461 (CausalStats—ERC-2017-ADG).

REFERENCES

ĆEVID, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2018). Spectral deconfounding and perturbed sparse linear models. Preprint. Available at arXiv:1811.05352.

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

DAWID, A. P. (2020). Decision-theoretic foundations for statistical causality. Preprint. Available at arXiv:2004.12493.

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

GUO, Z., ĆEVID, D. and BÜHLMANN, P. (2020). Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. Preprint. Available at arXiv:2004.03758.

JAKOBSEN, M. and PETERS, J. (2020). Distributional robustness of K-class estimators and the PULSE. Preprint. Available at arXiv:2005.03353.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S. et al. (2008). Genes mirror geography within Europe. *Nature* **456** 98–101.

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186 https://doi.org/10.1111/rssb.12167

PFISTER, N., WILLIAM, E., PETERS, J., AEBERSOLD, R. and BÜHLMANN, P. (2019). Stabilizing variable selection and regression. Preprint. Available at arXiv:1911.01850.

RICHARDSON, T. and ROBINS, J. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Preprint. Available at http://www.csss.washington.edu/Papers/wp128.pdf.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** 36. MR3862443

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229.

SHAH, R. D., FROT, B., THANEI, G.-A. and MEINSHAUSEN, N. (2018). RSVP-graphs: Fast high-dimensional covariance matrix estimation under latent confounding. *J. Roy. Statist. Soc. Ser. B.* Preprint. Available at arXiv:1811.01076.

THEIL, H. (1953). Repeated least squares applied to complete equation systems. The Hague: Central Planning Bureau.

THEIL, H. (1958). *Economic Forecasts and Policy*. North-Holland, Amsterdam, Netherlands.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839