Causal Inference under Incentives: An Annotated Reading List

KEEGAN HARRIS
Carnegie Mellon University
and
VASILIS SYRGKANIS
Stanford University

We provide an overview of research on causal inference in the presence of strategic agents. Work in this area uses tools from econometrics, statistics, machine learning, and game theory to infer causal relationships between treatments and outcomes of interest when the treated individuals have an incentive to behave strategically.

Categories and Subject Descriptors: F.7.2 [**Theory of computation**]: Algorithmic game theory; K.4 [**Computing methodologies**]: Machine learning

General Terms: Causal Inference, Game Theory, Machine Learning

Additional Key Words and Phrases: Decision-making, Incentives

Learning causal relationships from data is an important task across a wide variety of domains ranging from healthcare and drug development, to online advertising and e-commerce. As a result, there has been much work in the literature on economics, statistics, computer science, and public policy on designing algorithms and methodologies for causal inference.

While most of the focus has been on questions which are statistical in nature, one must also take game-theoretic incentives into consideration when doing causal inference about strategic individuals who have a preference over the treatment they receive. For example, it may be hard to infer causal relationships in randomized control trials when there is non-compliance by participants in the study. More generally, causal learning may be difficult whenever individuals are free to self-select their own treatments and there is sufficient heterogeneity between individuals with different preferences. Even when compliance can be enforced, individuals may strategize by modifying the attributes they present to the causal inference process in order to be assigned a more desirable treatment.

This annotated reading list is intended to serve as a brief summary of work on causal inference in the presence of strategic agents. While this list is not comprehensive, we hope that it will be a useful starting point for members of the SIGecom community to learn more about this exciting research area at the intersection of causal inference, game theory, and machine learning.

The reading list is organized as follows: (1, 3) study non-compliance in randomized trials, (2-4) focus on instrumental variable methods, (4-6) consider incentive misalignment between the individual running the causal inference procedure and

Authors' addresses: keeganh@cmu.edu, vsyrgk@stanford.edu

the subjects of the procedure, (7,8) study cross-unit interference, and (9,10) are about synthetic control methods.

(1) [Robins 1998]: This paper provides an overview of methods to correct for non-compliance in randomized trials (i.e., non-adherence by trial participants to the treatment assignment protocol).

(2) [Angrist et al. 1996]: This seminal paper outlines the concept of instrumental variables (IVs) and describes how they can be used to estimate causal effects. An IV is a variable that affects the treatment variable but is unrelated to the outcome variable except through its effect on the treatment. IV methods leverage the fact that variation in IVs is independent of any confounding to estimate the causal effect of the treatment.

(3) [Ngo et al. 2021]: Unlike prior work on non-compliance in clinical trials, this work leverages tools from information design to reveal information about the effectiveness of the treatments in such a way that participants become incentivized to comply with the treatment recommendations over time.

(4) [Harris et al. 2022]: This paper studies the problem of making decisions about a population of strategic agents. The authors make the novel observation that the assessment rule deployed by the principal is a valid instrument, which allows them to apply standard methods for instrumental variable regression to learn causal relationships in the presence of strategic behavior.

(5) [Miller et al. 2020]: This paper considers the problem of strategic classification, where a principal makes decisions about a population of strategic agents. Given knowledge of the principal's deployed assessment rule, the agents may strategically modify their observable features in order to receive a more desirable assessment. The authors are the first to show that designing good incentives for agent improvement (i.e. encouraging strategizing in a way which actually benefits the agent) is at least as hard as orienting edges in the corresponding causal graph.

(6) [Wang et al. 2023]: Incentive misalignment between patients and providers may occur when average treated outcomes are used as quality metrics. Such misalignment is generally undesirable in healthcare domains, as it may lead to decreased patient welfare. To mitigate this issue, this work proposes an alternative quality metric, the total treatment effect, which accounts for counterfactual untreated outcomes. The authors show that rewarding the total treatment effect maximizes total patient welfare.

(7) [Wager and Xu 2021]: Motivated by applications such as ride-sharing and tuition subsidies, this work studies settings in which interventions on one unit may have effects on others (i.e., cross-unit interference). The authors focus on the problem of setting supply-side payments in a centralized marketplace. They use a mean-field modeling-based approach to model the cross-unit interference, and design a class of experimentation schemes which allow them to optimize payments without disturbing the market equilibrium.

(8) [Li et al. 2023]: Like [Wager and Xu 2021], this paper studies the effects of cross-unit interference, although the interference considered here comes from

congestion in a service system. As a result, the interference considered here is dynamic, in contrast to the static interference considered in the previous entry.

(9) [Abadie and Gardeazabal 2003]: This is the first paper on synthetic control methods (SCMs), a popular technique for estimating counterfactuals from panel data. In the SCM setup, there is a pre-intervention time period during which all units are under control, followed by a post-intervention time period when all units undergo exactly one intervention (either the treatment or control). Given a test unit (who was given the treatment) and a set of donor units (who remained under control), SCMs use the pre-treatment data to learn a relationship (usually linear or convex) between the test and donor units. This relationship is then extrapolated to the post-intervention time period in order to estimate the counterfactual trajectory for the test unit under control.

(10) [Ngo et al. 2023]: A common assumption in the literature on SCMs is that of "overlap": the outcomes for the test unit can be written as a combination (e.g., linear or convex) of the donor units. This work sheds light on this often overlooked assumption and shows that (i) when units select their own treatments and (ii) there is sufficient heterogeneity between units who prefer different treatments, then overlap does not hold. Like [Ngo et al. 2021], the authors use tools from information design and multi-armed bandits to incentivize units to explore different treatments in a way which ensures that the overlap condition will gradually become satisfied over time.

REFERENCES

Abadie, A. and Gardeazabal, J. 2003. The economic costs of conflict: A case study of the basque country. *American economic review 93,* 1, 113–132.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91,* 434, 444–455.

Harris, K., Ngo, D. D. T., Stapleton, L., Heidari, H., and Wu, S. 2022. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*. PMLR, 8502–8522.

Li, S., Johari, R., Kuang, X., and Wager, S. 2023. Experimenting under stochastic congestion. *arXiv preprint arXiv:2302.12093*.

Miller, J., Milli, S., and Hardt, M. 2020. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*. PMLR, 6917–6926.

Ngo, D., Harris, K., Agarwal, A., Syrgkanis, V., and Wu, Z. S. 2023. Incentive-aware synthetic control: Accurate counterfactual estimation via incentivized exploration. *arXiv preprint arXiv:2312.16307*.

Ngo, D. D. T., Stapleton, L., Syrgkanis, V., and Wu, S. 2021. Incentivizing compliance with algorithmic instruments. In *International Conference on Machine Learning*. PMLR, 8045–8055.

Robins, J. M. 1998. Correction for non-compliance in equivalence trials. *Statistics in medicine 17,* 3, 269–302.

Wager, S. and Xu, K. 2021. Experimenting in equilibrium. *Management Science 67,* 11, 6694–6715.

Wang, S., Bates, S., Aronow, P., and Jordan, M. I. 2023. Operationalizing counterfactual metrics: Incentives, ranking, and information asymmetry. *arXiv preprint arXiv:2305.14595*.