Research Statement

Adam Bouyamourn

I am a PhD candidate in political methodology in the Political Science department at UC Berkeley, where I am advised by Thad Dunning (chair), Kirk Bansak, Erin Hartman, Avi Feller, and Peng Ding. I completed my training across Berkeley's Computer Science, Statistics, and Economics departments, and am an active member of Berkeley's causal inference community.

I have two broad areas of work. In the first area, I am interested in applying ideas from machine learning and optimization to problems in causal inference with a view to improving the generalizability, robustness, and external validity of findings in political science. In the second, I use game theory to study problems in selective inference. Both of these areas are relevant to political science practice: cumulative learning requires both the development of new reliable tools for research practice, and that the studies we actually conduct are in fact representative of the phenomena we are interested in. Leamer's classic essay, "Taking the 'Con' out of Econometrics", is known for its role in spurring the widespread adoption of causal inference methods, but also emphasizes the importance of selective inference concerns in improving the trustworthiness and generalizability of social science research.

Causal inference, machine learning and optimization

My job market paper, "Where To Experiment?", uses recent developments in operations research to study the site selection problem as an optimization problem. The contributions are theoretical and empirical. On the theory side, I show that we can understand the site selection problem for the PATE as the problem of choosing a set of sites with mean covariate vector as close as possible to that of the observed population. For the CATE, I show that this becomes a distributional discrepancy minimization problem. On the empirical side, I introduce a novel ensemble method that uses Best Subsets as a base learner, and show that it performs better on average than existing methods across a number of data sets. I also show how the researcher can incorporate prior information about site variances, treatment effects, and welfare into the problem.

This project has spurred an interest in the connections between Distributionally-Robust Optimization (DRO) and causal inference, which is the focus of work in progress. DRO methods involve solving estimation problems under the hypothesis that observed samples are drawn from a population that is within some uncertainty radius of the sample actually observed. DRO methods have a natural interpretation in terms of external validity: a finding is externally valid if it can be confirmed for a range of adversarially constructed distributional choices. Using DRO tools in political science is therefore a promising direction of research: we can develop methods that incorporate some degree of uncertainty about how closely our given study sample represents our actual population of interest.

It is natural to apply DRO methods to the site selection problem: what if the subpopulation of sites we actually observe differs from the population if sites we are in fact interested in? We can adapt the selection procedure to account for this distributional uncertainty. DRO methods are also applicable to analyzing experiments when we are interested in generalizing from a given subject population to a broader set of target populations.

In 2023, I published a solo-authored paper "Why LLMs Hallucinate", at *EMNLP*, a major NLP conference in Computer Science, on the causes of LLM hallucinations [Bouyamourn, 2023]. *EMNLP*

publishes very few solo-authored papers, and even fewer that are solo-authored by graduate students. The main result is a causal identification result. Semantic information and causal information both have a structural invariance property: if there is really a causal effect, changing the treatment changes the outcome; words have meanings, so changing a word changes the meaning of a sentence. Heuristically: LLMs hallucinate because they do not learn structurally invariant representations of the information they contain in their training data: they are stuck at the level of correlation, not causation. Grounding solves the problem, because it imposes a kind of structural invariance on LLM outputs: a sentence can be output only if it is a paraphrase of something that has been verified. There is quite a bit more to say on this topic, but I do not currently intend to pursue it as an active area of research.

I briefly describe three other causal inference projects undertaken during the PhD.

"The Power of Prognosis", with Clara Bicalho and Thad Dunning, develops balance tests for natural experiments that incorporate prognostic covariate information [Bicalho et al., 2022]. The broad idea is that when covariates are informative, and there are no interactions between observed and unobserved covariates, using an outcome model to weight covariates will give us better type I and type II error control over tests to falsify the hypothesis that relevant covariates are approximately balanced, as would be observed under randomization.

My contribution to "Prospective Experiment", published at *ACM e-Energy* with coauthors in computer science, was to consider the problem of separating out the effect of a reinforcement learning agent's training period from its efficacy as a decision maker [Spangher et al., 2020]. These are generally confounded: if an RL agent performs well at a task, is that because it has been trained for a long time, or because it is a better agent? My solution to this was to propose a within-subjects design to compare the performance of two RL agents: by assessing performance at two periods for each agent, we can separate out the effect of training period on agent performance.

"Identifying Causal Effects Of Exercise On Irregular Heart Rhythm Events Using Wearable Device Data", written with Lauren Hannah while I was an intern at Apple Health AI, uses data from the Apple Heart Study to assess the effect of exercise on atrial fibrillation risk. A puzzle in the medical literature is that exercise appears to both increase and decrease the risk of atrial fibrillation. By examining data from the Apple Watch, and using generalized propensity score methods, I found that, while exercise is protective on average, shocks to exercise level conditional on average exercise level increase atrial fibrillation risk. This helps to make sense of the conflicting findings in the medical literature: just looking at average effects can sometimes disguise the structure of the true, underlying response surface.

Selective inference and game theory

In the second broad area, I am interested in using game theory to study research incentives and problems in selective inference. This line of research was inspired by the work of, and discussions with, Will Fithian, and is similar to work being done by Michael Jordan and collaborators on strategic hypothesis testing [Fithian et al., 2017, Bates et al., 2024].

A given statistical analysis may be conditioned on a selection event that we may or may not observe. 'Selection bias' is the most straightforward example of this, but any research practice that affects the conclusion that we would draw from a given empirical application can be understood as such a conditioning event. Game theory is a natural method for studying selection problems with

strategic behavior, and the decision-theoretic setup of a given inference problem can be stated in game-theoretic terms.

A solo-authored paper, forthcoming at *Research and Politics*, "Collusive and Adversarial Replication", studies a situation in which social alignment between researchers and prospective replicators leads to a form of tacit collusion: Researchers can get away with conducting bad research, because social ties discourage prospective replicators from conducting ex post review of their research. While the paper studies the case of explicit misreporting by a Researcher, the general argument applies to 'softer' examples of bad practices. The moral is that adversarial ex post review is better for scientific progress.

A paper in progress, "Explore-Exploit: The Trade-Off Between Internal and External Validity", with Tak-Huen Chau, uses game theory to study what is meant by 'conflict' between internal and external validity. The basic idea is that a researcher choosing a study design faces an *explore-exploit* problem: they face a tension between optimizing a design for the probability of rejecting a given hypothesis; or they can maximize the generalizability or robustness of a given finding. This can be understood as a tension between representativeness and generalizability, and is referred to as 'the price of robustness' in the operations research literature. Methods that are optimal for minimizing average-case risk are not optimal for minimizing worst-case risk, and vice versa: there is no free lunch.

As a practical example, we study the adoption of pre-analysis plans. A possible downside of pre-analysis plans is that they may induce researchers to more heavily weigh exploitation rather than exploration in their study designs, potentially harming the generalizability of findings. Pre-analysis plans may make researchers opt for more conservative study designs. The recommendation is for pre-analysis plans to incorporate information on the universe of potential experiments that could have been chosen, and to explain in a principled way why the proposed experiment was chosen from that universe.

A second paper in progress, "On Double-Dipping", formalizes the notion of a research practice that violates selective inference standards. This adapts an idea of Eberhardt [2010], that we can understand causal discovery as a sequential game played by a Researcher against Nature. Double-dipping, in my formulation, is when the Researcher intercedes in the game against Nature: the Researcher, by acting after nature but before they conduct the analysis, is essentially playing against themselves.

An important point is that selective inference concerns can be independent of the widespread adoption of good methods. How an academic community is composed – how it regulates itself, and how it incentivizes its members – can be just as important as its adoption of better statistical technologies for conducting research.

References

Stephen Bates, Michael I. Jordan, Michael Sklar, and Jake A. Soloff. Incentive-theoretic bayesian inference for collaborative science, 2024. URL https://arxiv.org/abs/2307.03748.

Clara Bicalho, Adam Bouyamourn, and Thad Dunning. Conditional balance tests: Increasing sensitivity and specificity with prognostic covariates, 2022. URL https://arxiv.org/abs/2205.10478.

Adam Bouyamourn. Why LLMs Hallucinate, and How to Get (Evidential) Closure: Perceptual, Intensional, and Extensional Learning for Faithful Natural Language Generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.192. URL https://aclanthology.org/2023.emnlp-main.192.

Frederick Eberhardt. Causal discovery as a game. In Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors, *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pages 87–96, Whistler, Canada, 12 Dec 2010. PMLR. URL https://proceedings.mlr.press/v6/eberhardt10a.html.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2017.

Lucas Spangher, Akash Gokul, Manan Khattar, Joseph Palakapilly, Akaash Tawade, Adam Bouyamourn, Alex Devonport, and Costas Spanos. Prospective experiment for reinforcement learning on demand response in a social game framework. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, e-Energy '20, page 438–444, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380096. doi: 10.1145/3396851.3402365. URL https://doi.org/10.1145/3396851.3402365.