Applied Causality: Learning from Multiple Environments
Statistics GR8101, Spring 2023
Mon 1:00PM - 3:00PM
Uris 332

David M. Blei
Columbia University

We will study applied causality from the perspective of multiple environments.

In statistics and machine learning (ML), the idea of an "environment of data" appears in many guises, and many datasets naturally involve multiple environments. As some examples:

- genetic data about cells, in different organisms
- students taking a standardized test, in different schools
- medical histories of patients, in different hospitals
- lawmakers' votes, in different sessions of Congress

How can we analyze data from multiple environments? What kinds of statistical questions can we answer? What kinds of causal questions can we answer?

To help, the idea of multiple environments (sometimes with different terminology) appears in many threads of statistics and ML methods. For example, it is important to invariant learning, hierarchical modeling, synthetic controls, empirical Bayes, causal discovery, and causal representation learning. In this seminar, we will study these ideas, draw connections between them, and employ them to solve problems about multi-environment data.

Prerequisites and requirements. The class is open to doctoral students who have taken Foundations of Graphical Models (STCS6701) or know the material from that course. Each student will complete a research project about multi-environment learning.

We will set the syllabus as we go. Below are some texts we may discuss.

Hierarchical modeling and empirical Bayes

- Gelman and Hill (2007)
- Efron (2019)

Causal Representation Learning

- Bengio et al. (2019)
- Schölkopf et al. (2021)
- Wang and Jordan (2021)

Invariance and Causality

- Peters et al. (2016)
- Arjovsky et al. (2019)
- Lu et al. (2021)
- Yin et al. (2021)

Causal Discovery

- Zheng et al. (2018)
- Brouillard et al. (2020)
- Lopez et al. (2022)

Synthetic Controls

- Abadie et al. (2010)
- Abadie (2021)
- Agarwal et al. (2021)
- Athey et al. (2021)
- Shi et al. (2022)

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Agarwal, A., Shah, D., and Shen, D. (2021). Synthetic interventions. *arXiv:2006.0769*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv:1907.02893*.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *arXiv:1710.10251*.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv:1901.10912*.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Neural Information Processing Systems*.

Efron, B. (2019). Bayes, oracle Bayes, and empirical Bayes. *Statistical Science*, 34(2):177–201.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Lopez, R., Hutter, J., Pritchard, J., and Regev, A. (2022). Large-scale differentiable causal discovery of factor graphs. *Neural Information Processing Systems*.

Lu, C., Wu, Y., Hernandez-Lobato, J., and Schölkopf, B. (2021). Nonlinear invariant risk minimization: A causal approach. *arxiv:2102.12353*.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning. *arXiv:2102.11107*.

Shi, C., Sridhar, D., Misra, V., and Blei, D. (2022). On the assumptions of synthetic control methods. In *Artificial Intelligence and Statistics*.

Wang, Y. and Jordan, M. (2021). Desiderata for representation learning: A causal perspective. *arXiv:2109.03795*.

Yin, M., Wang, Y., and Blei, D. (2021). Optimization-based causal estimation from heterogenous environments. *arXiv:2109.11990*.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. *arXiv:1803.01422*.