Computational Methods and Causal Analysis
(POLSCI 689L)

Lectures: Tu & Th at 3:05 pm to 4:20 pm
Location: Languages 114B
Lab: Occasionally on Fridays, from 3:05 pm to 4:20 pm
Location: Languages 114B
Instructor: Jiawei Fu (jiawei.fu@duke.edu)
Office Hour: [link], Gross Hall 294A or online
TA: Trung-Anh Nguyen
Office Hour: TBD

Course Description

This course provides a guided exploration of advanced topics in quantitative methods, with a particular emphasis on computational approaches and modern causal inference. Students will engage with cutting-edge techniques, including machine learning and deep learning for unstructured data (such as text and images), nonparametric and semiparametric estimation, as well as their integration into contemporary frameworks for causal mechanisms and inference, such as double/debiased machine learning and targeted maximum likelihood estimation. Course content will evolve with developments in the field and incorporate recent research from the methodological literature. The course emphasizes both rigorous theoretical foundations and vibe coding experience.

The causal inference framework clarifies the identification assumptions required to estimate causal effects. However, there is no single unified approach to estimation. In essence, causal estimation relies on conditional expectation functions, which can be viewed as prediction tasks. Traditional methods often impose linear models, which may be overly restrictive. In this course, we will instead study more flexible predictive methods. Specifically, we will cover nonparametric and semiparametric estimation from an econometric perspective, as well as machine learning and deep learning from a computational perspective. We will also examine their theoretical properties.

Once we obtain consistent estimates of these predictive quantities, they can be incorporated into causal estimators. However, this leads to a two-step estimation procedure, which complicates statistical inference. Moreover, first-stage methods are often nonparametric and may introduce bias into the second stage. We will explore how to obtain robust estimates and valid inference using the double/debiased machine learning (DML) framework and the targeted learning framework (TMLE).

Finally, we will cover emerging topics in causal inference, such as proximal causal inference, heterogeneous treatment effects, causal inference with latent outcomes, and causal inference using unstructured data (e.g., from large language models).

Requirements and Grading

Problem Sets (50%): The most effective way to encourage learning and deepen understanding of the material is through hands-on assignments. Accordingly, there will be several problem sets designed to reinforce key concepts and provide practical experience. Each problem set will typically include two types of questions: (1) Theory—these questions focus on applying formulas and concepts discussed in class; and (2) Programming—these tasks involve vibe coding to implement the computational methods covered in the course.

Midterm (25%): Computational Causal Analysis Competition. I will provide a dataset for you to estimate treatment effects and heterogeneous treatment effects. You may use any model of your choice. Your grade will be based on the performance of your estimates.

Final Project (25%)

Please choose one of the following three options for your project:

(1) Apply the Method in Your Own Ongoing Research Project. The ultimate goal of methods training is practical application. If you are currently working on a research project, you can incorporate the method learned in class or related into your project. This can be in your main text or in an appendix. Doing so not only helps you practice the method but also advances your own research—truly a win-win solution.

(2) Replication and Extension. Select an applied paper of interest that uses one of the methods introduced in class or related. Replicate its main findings to demonstrate your understanding. In the final section of your replication report, add a new contribution and implement it — such as a meaningful extension, areas for improvement, etc,

(3) Methodological Proposal. If you are more interested in the methodology itself, use this assignment to propose a new research project focusing on the method. Your proposal should include: a) A clear research objective (e.g., solving a puzzle, answering a question, or introducing a new method); b) A literature review; c) Some preliminary work (e.g., outlining a conceptual framework, presenting conjectured results, running a simple simulation, or proposing a possible approach).

AI and Collaboration Policy

AI and your peers are valuable resources for study and research. In this course, we will actively learn how to use AI and incorporate it into our own research workflows. However, these tools are most effective when you have built a strong foundation of knowledge yourself. For example, it is well known that generative AI can produce incorrect or misleading information (often referred to as "hallucinations"). As a user, you can identify such errors only if you have a solid understanding of the subject. Taking courses, studying, and practicing are therefore essential for developing the foundational skills needed to engage with AI effectively.

Accordingly, for theory problems, please do not rely on AI to solve them for you. Simply reading AI-generated answers will not support meaningful learning. Real understanding develops through careful thinking and working through challenges on your own.

For programming problems, you are encouraged to use AI tools (e.g., for coding assistance). However, you remain fully responsible for the correctness of your code and the validity of your results.

You are also encouraged to discuss ideas with your peers, but all submitted work must be written independently. For each assignment, please clearly indicate: (i) which parts, if any, were generated with AI assistance, and (ii) with whom you discussed the assignment. Honest documentation of your learning process is essential for developing genuine understanding and maintaining academic integrity.

Resources

We will not strictly follow a specific textbook. Instead, we will draw from various sources from following references.

- Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.
- Wasserman, L. (2006). All of nonparametric statistics. Springer Science & Business Media.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, No. 1). New York: springer.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Fan, J., Li, R., Zhang, C. H., & Zou, H. (2020). Statistical foundations of data science. Chapman and Hall/CRC.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
- Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.
- Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge university press.
- Ding, P., 2024. A first course in causal inference. Chapman and Hall/CRC.
- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. arXiv preprint arXiv:2403.02467.
- Wager, S. (2024, September). Causal inference: A statistical learning approach.
- Jurafsky, D. and Martin, J. H., (2025). Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Dive into deep learning. Cambridge University Press.
- Bishop, C. M., & Bishop, H. (2023). Deep learning: Foundations and concepts. Springer Nature.
- Prince, S. J. (2023). Understanding deep learning. MIT press.
- Van der Laan, M.J. and Rose, S., 2018. Targeted learning in data science. Cham: Springer International Publishing.

- Van der Laan, M.J. and Rose, S., 2011. Targeted learning: causal inference for observational and experimental data (Vol. 4). New York: Springer.

Topics

Topic 1: Introduction to Computational Causal Analysis

- Overview of the course
- Develop an understanding of how and why computational methods contribute to causal inference

Topic 2: Linear ML Method

- Penalized Least-Squares
- LASSO and Adaptice LASSO
- Elastic Net
- Ridge

Topic 3: Tree-based Methods and Boosting

- Decision Trees
- Boosting
- Random Forests

Topic 4: Nonparametric Regression

- Kernel Regression
- Local linear and Polynomial Estimators
- Asymptotic Integrated MSE
- Bandwith Selection
- Polynomial Basis
- Splines
- Large Sample Properties

Topic 5: ML and Causal Inference I

- Doubly Robust Estimator
- Neyman Orthogonality
- Double LASSO
- Double/De-biased machine learning
- DML for IV, DID, RDD

Topic 6: ML and Causal Inference II

- Heterogeneous Treatment Effects
- Meta-Learners
- BLP tests
- Generalized random forest

Topic 7: Deep Learning I: Neural Networks

- Universal approximation theorem
- Double descent

Topic 8: Deep Learning II: More Architectures

- CNNs,GNNs
- Transformer
- LLM

Topic 9: Unstructured Data: Text data

- Descriptive Inference
- Similarity,Complexity,Sentiment
- Structural Topic Models
- LDA
- Bert

Topic 10: Causal Inference with Unstructured Data

- Proximal causal inference
- Text/Figure/Video as outcome, treatment
- Latent outcome
- Latent confounder

Topic 11: Semiparametric Inference and TMLE

- Semiparametric Efficiency Theory
- One-step Estimator
- TMLE

Topic 12: Your Interested Topics