Computational Methods in Social Science

(POLSCI 690-5 Advanced Topics in Political Methodology)

Lectures: Tu & Th at 3:05 pm to 4:20 pm
Location: Gross Hall 111
Instructor: Jiawei Fu (jiawei.fu@duke.edu)
Office Hour: [link], Gross Hall 294A or online

Course Description

This course provides a guided exploration of advanced topics in quantitative methods, with a special focus on computational methods, aiming to reach the current frontiers of the field. Students will engage with cutting-edge techniques, including machine learning, deep learning, social network analysis, and data-driven approaches to text and image analysis. They will apply these methods in their own empirical research or work on developing methodological innovations.

We will focus on four main topics. (1) High-dimensional statistics and Learning: With the rise of computational power, statistical learning has become a powerful tool for flexibly modeling underlying functions and making predictions with high-dimensional data. We will explore these techniques and their applications, including recent advancements in using machine learning for experimental design and causal inference, such as detect HTE and double machine learning. (2) Text Analysis: This section applies machine learning to analyze textual data, doing causal inference with unstructured data, providing students with tools to extract meaningful insights from large-scale text sources. (3) Deep learning: As a type of machine learning that focuses on neural networks, it has become one of the most popular techniques today. We will cover the basic concepts and explore applications in image analysis and natural language processing. (4) Social network analysis: We will introduce the background and methods for measuring and representing networks, followed by two modeling approaches: the random graph model and the game-theoretical strategic model.

This course emphasizes both theoretical foundations and hands-on programming experience with real data. A working knowledge of statistical inference, linear algebra, calculus, elementary econometrics, and R programming is assumed.

Requirements and Grading

Problem Sets (50%): The most effective way to encourage learning and deepen understanding of the material is through hands-on assignments. Consequently, there will be some problem sets designed to reinforce key concepts and provide practical experience. Each problem set will typically include two types of questions: 1) Simple algebra: These questions only require the direct application of formulas discussed in class; 2) Programming: These tasks involve using R to write code and implement computational methods covered in the course.

Midterm (25%): Machine Learning Competition. I will send out a training dataset for you to learn at home. You may train any model you like. On the day of the midterm, I will provide a test dataset in class. You will then use your trained model to make predictions. Your grade will be based on the performance of your model.

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 social science paper of interest that uses one of the methods introduced in class or related. Replicate its main findings in R Markdown to demonstrate your understanding. In the final section of your replication report, add a new contribution and implement it in R — 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. However, they 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 (hallucinations). As a user, you can only identify these mistakes if you have a solid understanding of the subject. Taking classes, studying, and practicing are essential steps to develop the foundational skills needed to effectively engage with AI in the future.

Therefore, please do not upload your homework and rely on AI to solve problems for you. Simply reading AI-generated answers will not help you learn. Real learning happens when you think deeply, and struggle through challenges.

You are encouraged to discuss with your peers, but the final work must be written by you. For any submitted work, please indicate: which parts, if any, were generated with AI assistance, and who you discussed the assignment with. Honest documentation of your learning process ensures that you develop real understanding and academic integrity.

Resources

We will not strictly follow a specific textbook. Instead, we will draw from various sources. Below are some excellent general references that may be helpful.

- 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.
- 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.
- Jackson, M. O. (2008). Social and economic networks. Princeton: Princeton university press.
- Goyal, S., 2023. Networks: An economics approach. MIT Press.

Schedule

Week 1 (Aug 26-28): Introduction to High-dimensional Statistics and Learning

- Overview of the course
- Supervised and unsupervised learning
- Bias-Variance Trade-Off

Week 1-2 (Aug 28-Sep 2): Linear Methods and Regularization

- Ridge regression
- LASSO
- Cross validation

Week 2 (Sep 4) Post-selection Inference

- Selective Inference
- Simultaneous Inference

Week 3 (Sep 9-11): Tree-based Methods and Boosting

- Decision Trees
- Boosting
- Random Forests

Week 4 (Sep 11): Unsupervised learning

- K-means
- PCA

Week 5 (Sep 16-23): ML and Causal Inference I

- Heterogeneous Treatment Effects
- S-leaner, L-learner, X-learner
- Generalized random forest

Week 6 (Sep 23- Qct 7): ML and Causal Inference II

- Double/De-biased machine learning
- Covariate Selection
- Identification strategies with ML

Week 7 (Oct 7-9): Interested Topics, and Midterm (Oct 7)

Tell me your interested topics.

Oct 14: Fall Break

Week 8 (Oct 16-21): Text as Data I: Descriptive Inference and Classification

- Similarity
- Complexity
- Naive Bayes and SVM
- Sentiment

Week 9 (Oct 23): Text as Data II: Topic Model

- Structural Topic Models
- LDA
- Bert

Week 10 (Oct 28): Text as Data III: Causal Inference

- Latent representation
- Proximal causal inference

Week 11 (Oct 30): Deep Learning I: Neural Networks and Variant Architectures

- Universal approximation theorem
- Double descent
- CNNs,GNNs
- Transformer

Week 12 (Nov 4-6): Deep Learning II: Representation Learning and Causal Inference

- Representation Learning
- Causal Learning

Week 13 (Nov 11-13): Social Network Analysis I

- Representing and measure network
- Random-Graph model
- Strategic network formation

Week 13 (Nov 18-20): Social Network Analysis II

- Causal Inference
- Application

Week 14 (Nov 25-27): Thanksgiving Recess

Dec 15: Final Project Due