17.806: Quantitative Research Methods IV

Spring 2021

Instructor: In Song Kim
TA: Tomoya Sasaki

Department of Political Science
MIT

1 Contact Information

	In Song	Tomoya
Office:	E53–407	E53–458
Email:	insong@mit.edu	tomoyas@mit.edu
Phone:	617–253–3138	
URL:	http://web.mit.edu/insong/www	https://polisci.mit.edu/people/tomoya-sasaki

2 Logistics

- Lectures: Tuesdays and Thursdays, 3pm–4:30pm
- Recitations: Friday, 9–10am

Note that the first class meets on February 16. No class will be held on March 9 (Monday schedule), March 23, and April 20 (Student holidays). Last day of class is May 20.

3 Course Description

This course is the fourth and final course in the quantitative methods sequence at the MIT political science department. The course covers various advanced topics in applied statistics, including those that have only recently been developed in the methodological literature and are yet to be widely applied in political science. The topics for this year are organized into three broad areas: (1) research computing, where we introduce various techniques for automated data collection, visualization, and analysis of massive datasets; (2) statistical learning, where we provide an overview of machine learning algorithms for predictive and descriptive inference as well as their applications in causal inference methods; and (3) finite mixture models (e.g., Latent Dirichlet allocation for text analysis), as well as a variety of estimation techniques such as the EM algorithm and Variational Inference.

4 Prerequisites

There are three prerequisites for this course:

1. Mathematics: multivariate calculus and linear algebra.

2. Probability and statistics covered in 17.800, 17.802 and 17.804, including linear regression, causal inference, and Bayesian statistics.

3. Statistical computing: proficiency with at least one statistical software. We will use `R` in this course (more on this below).

For 1, refer to this year's math camp materials to see the minimum you need to know; see

Math Camp 1: `https://stellar.mit.edu/S/project/mathprefresher/`
Math Camp 2: `https://canvas.mit.edu/courses/5733`

This class will assume that you have already had some prior exposure to the material covered and go through many concepts relatively quickly.

5 Course Requirements

Emergency academic procedures are in effect in Spring 2021 due to the COVID-19 outbreak. The final grades are based on the following items. All students may elect to have one subject graded PE/NE in the spring semester instead of the letter grades (A, B, C, D/NE, and F/NE). For further information, please check this document.

- **Problem sets** (45%): **Six** bi-weekly problem sets will be given throughout the semester. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will contribute equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted.

  - All answers should be typed. Students are strongly encouraged to use LaTeX, a typesetting system that has become popular in the field (or LaTeX typesetting in RMarkdown). Please make sure that your code follows the Google and tidyverse `R` style guide rules (URLs are here and here).
  - Late submission will not be accepted unless you ask for special permission from the instructor in advance (Permission may be granted or not granted, with or without penalty, depending on the specific circumstances).
  - Working in groups is encouraged, but each student must submit their own writeup of the solutions. In particular, you should not copy someone else's answers or computer code. We also ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.
  - For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For data analysis questions, include annotated code as part of your answers. All results should be presented in a single document so that they can be easily understood. RMarkdown is strongly encouraged.

- **Final project** (50%): The final project will be a paper which applies methods learned in this course to an empirical problem of your substantive interest.

1. **Data and Initial Analysis** (15%)
   - Students are expected to collect their own data related to an empirical problem of own interest.
   - Students who do not have particular target data sources should consult with the instructor by March 4.
   - Replication papers are allowed, but you must go beyond the original analysis in some significant way by collecting additional data *and* applying techniques learned in the course. If you have any doubts, please consult with the instructor and TA.
   - March 25: A one-page memo due (see below).
   - April 22: A five-page report due (see below).
2. **Paper** (35%): The paper should be *maximum 10 pages* of double-spaced 12-point font text (including references and appendix) with 1-inch margins.
   - Title
   - Abstract (150 words)
   - Introduction: Introduction must contain the following.
     (a) The problem/puzzle to be solved
     (b) Explain why previous work and methods leave the problem unresolved
     (c) Your contribution, i.e., the solution to the problem/puzzle. You need to give the reader a clear sense of how you will solve the problem.
     (d) Brief summary of your findings
   - Data section: Describe your novel data collection efforts
   - Empirical analysis: Figures and tables with informative captions

Collaboration: We encourage you to collaborate with another student (a group should not consist of more than 2 students). Note that most cutting-edge research is collaborative (see any recent issue of *APSR* or *AJPS*), and collaboration is more likely result in a good, potentially publishable paper (multiple brains are usually better than one).

Please be aware of the following deadlines. Late submission will be penalized.

- **March 25 (Data Collection):** By this date, you should acquire the data to be analyzed and start preliminary descriptive data analysis. Please upload one-page memo to the Canvas website with the following components.
  * Main theoretical/empirical contributions/motivations
  * Data description (source, collection methods, and why better than previous data)
- **April 22 (Initial descriptive analysis):** By this date, you should submit a five-page report summarizing your data collection and descriptive data analysis to the Canvas website.
- **May 20 (Final Paper)**: By this date, you should submit your final paper to the Canvas website by midnight.

• **Participation** (5%): Students are strongly encouraged to ask questions and actively participate in discussions during lectures and recitation sessions. In addition, there will be recommended readings for each section of the course which students are strongly encouraged to complete prior to the lectures in order to get the most out of them.

6 Course Website

You can find the Stellar website for this course at:

`https://canvas.mit.edu/courses/6765`

We will distribute course materials, including readings, lecture slides, and problem sets, on this website.

7 Questions about Course Materials

In addition to recitation sessions and office hours, please use the "Course Question Board" under the Canvas discussion page when asking questions about lectures, problem sets, and other course materials.

Using the Canvas will allow students to see other students' questions and learn from them. Both the TA and the instructor will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. A student's respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructor or TA* (unless they are of a personal nature)—we will not answer them!

8 Recitation Sessions

Weekly recitation sessions will be held at Zoom on Fridays. Sessions will cover a review of the theoretical material and also provide help with computing issues. The teaching assistant will run the sessions and can give more details. Attendance is strongly encouraged.

9 Notes on Auditing

In order to audit this course, one must

- Obtain the course instructor's permission
- Complete all problem sets
- Submit written comments on each project's descriptive data analysis

10 Notes on Computing

- In this course we use `R`, an open-source statistical computing environment that is very widely used in statistics and political science. (If you are already well versed in another statistical software, you are free to use it, but you will be on your own.) Each problem set will contain computing and/or data analysis exercises which can be solved with `R` but often require going beyond canned functions to write your own program. We provide problem set solutions using `R`.
- We strongly encourage you to use RMarkdown. These are useful resources to learn about RMarkdown
  - Tierney, Nicholas. *RMarkdown for Scientists* [Link].

- Xie, Yihui, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook* [Link].
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide* [Link].

- Following reference would be useful to write clean and efficient code in `R`
  - Google's style guide [Link].
  - Tidyverse style guide [Link] (You do not need to use the Tidyverse but chapters 1–3 are very useful for non-Tidyverse users as well).

- If your project requires large computational resources, we recommend using xvii or Research Computing Environment (RCE) available through the Harvard-MIT Data Center (HMDC).

11 Books

- Recommended books: We will read chapters from these books throughout the course. We strongly recommend that you at least purchase Bishop. These books will be available for purchase at COOP and online bookstores (e.g. Amazon) and on reserve in the library.
  - Bishop, Christopher M. 2007. *Pattern Recognition and Machine Learning.* Springer (a great introduction to machine learning).
  - Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning.* Springer.
  - Murphy, Kevin P. 2012. *Machine Learning.* The MIT Press.
  - James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning.* Springer.
  - Bühlmann, Peter and Sara van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.
  - Jurafsky, Daniel and James Martin. 2018. *Speech and Language Processing.* Prentice Hall. PDF

12 Course Outline

12.1 Introduction

1. Big Data in Political Science

   Recommended Reading:

   - Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.
   - Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: an Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.
   - Athey, Susan and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics*, 11, 685–725.
   - Grimmer, Justin. 2015 "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48 (1): 80–83.

12.2 Automated Data Collection

1. Web Scraping, Regular Expressions

   Recommended Reading:

   - Jurafsky and Martin 2.1.
   - For a basic tutorial on HTML, consult 3 sources linked from this blog post: Three great places to start learning HTML.
   - Jackman, Simon. 2006. "Data from the Web Into `R`" *The Political Methodologist.* 14 (2): 11–15.
   - Data Camp Course: Working with Web Data in R.

12.3 Supervised Learning

1. Suppor Vector Machine (SVM)

   Recommended Reading:

   - Bishop Appendix E. Lagrange Multipliers.
   - Bishop 7.1 (7.1.3, 7.1.4 optional).
   - Murphy Ch.14 (optional).
   - Bonica, Adam. 2018. "Inferring Roll–Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science* 62 (4): 830–848.

2. Over-fitting (Model Selection), Cross-validation

   Required Reading:

   - Hastie, Tibshirani, and Friedman Ch.7.

   Recommended Reading:

   - Bishop 1.1.

3. Variable Selection (Ridge Regression, LASSO)

   Required Reading:

   - Hastie, Tibshirani, and Friedman 3.1–3.4.

   Recommended Reading:

   - Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu. 2016. "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments." *Proceedings of the National Academy of Sciences* 113 (27): 7383–7390.
   - Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

4. Additive Models & Ensemble Methods: Generalized Additive Models (GAM), Bagging, Boosting, Random Forests

Recommended Reading:

- Hastie, Tibshirani, and Friedman Chs.9, 15, 16.
- Montgomery, Jacob and Santiago Olivell. "Tree-Based Models for Political Science Data." 2018. *American Journal of Political Science* 62 (3): 729–744.
- Bishop Ch.14.
- Murphy Ch.16.

12.4 Machine Learning for Causal Inference

1. Machine learning for Causal Inference

   Required Reading:

   - Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.
   - Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–4165.

   Recommended Reading:

   - Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–434.
   - Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2017. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *American Economic Review* 107 (5): 261–265.
   - Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (4): 597–623.
   - Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.
   - Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.
   - Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7 (1): 443–470.

12.5 Dimension Reduction

1. Principal Component Analysis, Factor Analysis

   Recommended Reading:

   - Bishop Ch.12 (towards 12.2.1).
   - Hastie, Tibshirani, and Friedman 14.5.

- Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109 (1): 62–78.
- Heckman, James J. and James M. Snyder. (1997). "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators." *RAND Journal of Economics* 28: S142–S189.
- Treler, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52 (1): 201–217.
- Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77 (4): 1229–1279.

2. T-SNE

   Recommended Reading:

   - L.J.P. van der Maaten and G.E. Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (11): 2579–2605.

12.6 Mixture Models

1. Probability Distributions

   Required Reading:

   - Bishop 2, Appendix B.

2. EM Algorithm

   Required Reading:

   - Bishop Ch.9.

   Recommended Reading:

   - Murphy 11
   - Imai, Kosuke, and Dustin Tingley. 2012. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56 (1): 218–236.
   - Jackman, Simon. 2000. "Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo." *American Journal of Political Science* 44 (April): 375–404.

3. Variational Inference

   Required Reading:

   - Grimmer, Justin. 2010. "An introduction to Bayesian inference via variational approximations." *Political Analysis* 19 (1): 32–47.

   Recommended Reading:

- Bishop Ch.10.
- Murphy Ch.21.
- Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110(4): 631–656.

12.7 Text Analysis

1. Text as Data: regular expression, stemming

   Recommended Reading:

   - Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.
   - Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." *Journal of Economic Literature* 57(3): 535–74.
   - Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26 (2): 168–189.

2. Topic models: Latent Dirichlet Analysis, Correlated Topic Models, Structural Topic Models

   Recommended Reading:

   - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3: 993–1022.
   - Blei, David, and John Lafferty. 2006. "Correlated Topic Models." *Advances in Neural Information Processing Systems* 18: 147.
   - Roberts, Margaret E., Stewart Brandon M., and Airoldi Edo M. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003.
   - Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082.

3. Words and Votes: Scaling with Text

   Recommended Reading:

   - Gerrish, Sean, and David M. Blei. 2012. "How they vote: Issue-adjusted models of legislative behavior." *Advances in Neural Information Processing Systems*.
   - Lauderdale, Benjamin E., and Tom S. Clark. 2014. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58 (3): 754–771.
   - Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–722.

- Kim, In Song, John Londregan, and Marc Ratkovic. 2018. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* 26 (2): 210–229.

4. Word Embeddings

   Recommended Reading:

   - Jurafsky and Martin Ch.6.
   - Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." URL: `https://arxiv.org/abs/1301.3781`
   - Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." URL: `https://arxiv.org/abs/1310.4546`
   - Rheault, Ludovic and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora" *Political Analysis* 28 (1): 112–133.

12.8 Causal Inference with Time-Series Cross-Section Data

Recommended Reading:

- Imai, Kosuke and In Song Kim. 2019. "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63 (2): 467–490.
- Imai, Kosuke, In Song Kim, and Erik Wang. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." Working paper available at `http://web.mit.edu/insong/www/pdf/tscs.pdf`
- Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25 (1) 57–76.
- Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. "Matrix Completion Methods for Causal Panel Data Models." `https://arxiv.org/abs/1710.10251`.
- De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.
- Imai, Kosuke, and In Song Kim. 2020. "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*. Forthcoming.

12.9 Network Models (Time Permitting)

Recommended Reading:

- Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. 2002. "Latent Space Approaches to Social Network analysis." Journal of the American Statistical Association 97, (460): 1090–1098.

- Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. "Mixed Membership Stochastic Blockmodels." *Journal of Machine Learning Research.*
- Barberá, Pablo. 2014. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data", *Political analysis* 23 (1): 76–91.
- Kim, In Song and Dmitriy Kunisky. 2020. "Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics." *Political Analysis.* Forthcoming. `http://web.mit.edu/insong/www/pdf/network.pdf`.
- Cranmer, Skyler J and Desmarais, Bruce A. 2011."Inferential Network Analysis with Exponential Random Graph Models". 2011. *Political Analysis.* 19 (1): 66–86.
- Egami, Naoki. 2020. "Spillover Effects in the Presence of Unobserved Networks" *Political Analysis.* Forthcoming.