17.806: Quantitative Research Methods IV

Spring 2018

Instructor: In Song Kim
TA: Andy Halterman

Department of Political Science
MIT

1 Contact Information

	In Song	Andy
Office:	E53–407	E40–446
Email:	insong@mit.edu	ahalt@mit.edu
Phone:	617–253–3138	617-866-0547
URL:	http://web.mit.edu/insong/www	https://polisci.mit.edu/people/andrew-halterman

2 Logistics

- Lectures: Mondays and Wednesdays 3:30–5:00pm, E53–438
- Recitations: Friday, 9am–10am, E53–438
- In Song's office hours: Thursday 4:00pm–6:00pm
- Andy's office hours: Monday 11am–1pm, E40-446.

Note that the first class meets on February 7. We will hold class on Tuesday, February 20 (Monday schedule). No class will be held on April 16 (Patriots Day). Last day of class is May 16.

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; 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, Bayesian statistics
3. Statistical computing: proficiency with at least one statistical software. We will use `R` in this course (more on this below).
4. Python workshop: All students are required to attend the two Python workshops offered in February: "Introduction to Data Science with Python", February 9, 2 4 p.m and "Web Scraping with Python" February 23, 2 4 p.m. .

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://stellar.mit.edu/S/project/mathcamp2/

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

The final grades are based on the following items:

- **Problem sets** (45%): **Seven** 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. Please make sure that your code follows Google's `R` Style Guide rules (here is the URL).
  - Neither late submission nor electronic submission will 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** (10%)
   - Students are expected to collect their own data related to an empirical problem of own interest. Attending the Political Methodology Lab Workshops is strongly encouraged. The first two are required: (1) Introduction to Data Science with Python (February 9, 2–4 p.m), and (2) Web Scraping with Python (February 23, 2–4 p.m).
   - Students who do not have particular target data sources should consult with the instructor by February 16.
   - 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.
2. **Paper** (35%)
   - Title
   - Abstract (150 words)
   - Introduction (2 pages max): 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 (2 pages max)
   - Figures and tables with informative captions
3. **Presentation** (5%) Students will give presentations in front of the class during the regular class time (early May). Presentations should last about 10 minutes (determined based on the class size, but time limits will be strictly enforced).

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).

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

- **March 23 (Descriptive data analysis):** By this date, you should acquire the data to be analyzed and conduct preliminary descriptive data analysis. Please upload a brief memo to the Stellar webpage with the following components.
  * Main theoretical/empirical contributions/motivations
  * Data description (why better than previous data)
  * Up to three Figures/tables with informative captions
- **April 29 (Initial analysis):** By this date, you should finish initial data analysis. Meet with the instructor to get feedback on your analysis (schedule a meeting with the instructor in the week of April 16).
- **May 16 (Final Paper)**: By this date, you should submit your final paper to the Stellar webpage by midnight.

  - **May 18 (Poster Presentation) TBD**

- **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:

http://stellar.mit.edu/S/course/17/sp18/17.806/

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 Piazza Q&A board when asking questions about lectures, problem sets, and other course materials. You can access the Piazza course page either directly from the below address or the link posted on the Stellar course website:

https://piazza.com/mit/spring2018/17806

Using Piazza 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 in E53-438 on Fridays 10:30–11:30am. 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
- Register officially as a listener
- Complete all problem sets
- Submit comments on each project's descriptive data analysis by April 5.

10 Notes on Poster

Poster presentation is an efficient way to get valuable feedback from a large number of people. A poster should follow the structure of your paper, and thus it is a helpful way to think about the organization of your paper before writing it. Here are some notes.

1. **Use keywords and bullet points:** You should not use full sentences—your audience will never read them. Try to use keywords (or half sentences when needed), and make sure that you use only one line to deliver each point.
2. **Use LaTeX:** There are many online templates to help you make posters easily, e.g., http://www-i6.informatik.rwth-aachen.de/ dreuw/latexbeamerposter.php
3. **Examples:** You may find it helpful to look at some of the posters presented at Political Methods conferences. It's available here.

11 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.
- If your project requires large computational resources, I recommend using Research Computing Environment (RCE) available through the Harvard-MIT Data Center (HMDC).

12 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.
  - Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning*, Springer (A great introduction to machine learning).
  - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.
  - Kevin P. Murphy. 2012. *Machine Learning*, The MIT Press
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014 *An Introduction to Statistical Learning.* Springer.
  - Bühlmann P, Van De Geer S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 2011. Springer.

13 Tentative Course Outline

13.1 Introduction

1. Big Data in Political Science

   Recommended Reading:

- Burt L. Monroe, Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2015. " No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science & Politics* 48 (1).
- Hal R. Varian. 2014. "Big Data: New Tricks for Econometrics" *Journal of Economic Perspectives* 28(2), 3–28

2. Rcpp, Armadillo for research computing

13.2 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)

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

   Required Reading:

   - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 7.

   Recommended Reading:

   - Bishop 1.1
   - Kass, Robert E. and Adrian E. Raftery. "Bayes Factors." *Journal of the American Statistical Association*, 29, 773–795.

3. Variable Selection (Ridge Regression, LASSO)

   Required Reading:

   - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 3.1–3.4

   Recommended Reading:

   - Tibshirani, Robert. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
   - Bühlmann, Peter, and Sara Van De Geer (2011). "Statistics for high-dimensional data: methods, theory and applications." *Springer Science & Business Media*
   - Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon, Bin Yu 2016) "Lasso adjustments of treatment effect estimates in randomized experiments." *PNAS* (forthcoming).

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

   Recommended Reading:

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 9, 15, 16
- Bishop 14
- Murphy Ch.16

5. Machine learning for Causal Inference

   Required Reading:

   - Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 29–50

   Recommended Reading:

   - Susan Athey and Guido Imbens. 2016. "Recursive Partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360
   - Stefan Wager and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." Forthcoming *Journal of the American Statistical Association*.
   - Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2017. "Double/Debiased Machine Learning for Treatment and Structural Parameters." Working paper available at `https://arxiv.org/pdf/1608.00060.pdf`

13.3 Dimension Reduction

1. Principal Component Analysis

   Recommended Reading:

   - Bishop Ch. 12 (towards 12.2.1)
   - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 14.5

2. Factor Analysis

   Recommended Reading:

   - 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*: S142–189
   - Shawn Treler and Simon Jackman, "Democracy as a Latent Variable." *American Journal of Political Science*: 201–217

13.4 Mixture Models

1. Probability Distributions

   Required Reading:

   - Bishop Ch.2, Appendix B

2. EM Algorithm

   Required Reading:

   - Bishop Ch.9

   Recommended Reading:

   - Murphy Ch.11
   - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
   - Jackman, Simon. "Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo." *American Journal of Political Science* (2000): 375-404.

3. Variational Inference

   Required Reading:

   - Grimmer, Justin. "An introduction to Bayesian inference via variational approximations." *Political Analysis* (2010)

   Recommended Reading:

   - Bishop Ch.10
   - Murphy Ch.21
   - Kosuke Imai, James Lo, and Jonathan Olmsted. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110(4), 631–656. (2016) (2016)
   - Ryan J. Giodano, Tamara Broderick, and Michael I. Jordan. "Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes." Advances in Neural Information Processing Systems (2015)
     - See here for an implementation

13.5 Text Analysis

1. Text as Data: regular expression, stemming

   Recommended Reading:

   - Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* (2013): 28.
   - Spirling, Arthur. "US treaty making with American Indians: Institutional change and relative power, 17841911." *American Journal of Political Science* 56.1 (2012): 84-97.

- Quinn, Kevin M., et al. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54.1 (2010): 209-228.

2. Latent Dirichlet Analysis

   Recommended Reading:

   - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3 (2003): 993-1022.
   - Taddy, Matt. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association* 108.503 (2013): 755-770.

3. Correlated Topic Models

   Recommended Reading:

   - Blei, David, and John Lafferty. "Correlated topic models." *Advances in Neural Information Processing Systems* 18 (2006): 147.

4. Structural Topic Models

   Recommended Reading:

   - Roberts Margaret E, Stewart Brandon M, Airoldi Edo M. "A model of text for experimentation in the social sciences."' *Working Paper* (2015).
   - Roberts, Margaret E., et al. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* (2014).

5. Words and Votes: Scaling with Text

   Recommended Reading:

   - Gerrish, Sean, and David M. Blei. "How they vote: Issue-adjusted models of legislative behavior." *Advances in Neural Information Processing Systems*. 2012.
   - Lauderdale, Benjamin E., and Tom S. Clark. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58.3 (2014): 754-771.
   - Slapin, Jonathan B., and Sven-Oliver Proksch. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52.3 (2008): 705-722.
   - Kim, In Song, John Londregan, and Marc Ratkovic. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* Forthcoming. 2018

13.6 Network Models

Recommended Reading:

- Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed Membership Stochastic Blockmodels." *Journal of Machine Learning Research*. 2008.
- Kim, In Song and Dmitriy Kunisky. "Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics." Working paper available at http://web.mit.edu/insong/www/pdf/network.pdf, 2018

13.7 Causal Inference with Time-Series Cross-Section Data

Recommended Reading:

- Imai, Kosuke and In Song Kim. "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Working paper available at http://web.mit.edu/insong/www/pdf/FEmatch.pdf. 2017

13.8 Sequential Data

1. Hidden Markov Models

Recommended Reading:

- Bishop Ch.13
- Murphy Ch.17
- Park, Jong Hee. "A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models." *American Journal of Political Science* 56.4 (2012): 1040-1054.