GOV 2017: Applied Bayesian Statistics for the Social Sciences

Kosuke Imai

Spring 2023

Preliminary syllabus

Abstract

This course introduces social science students to applied Bayesian statistics. We will begin by introducing Bayes' rule, which allows us to learn from data in an intuitive and coherent way. We then cover a set of simple probabilistic models as well as powerful computational tools that will be used for the remainder of the course. Finally, we will learn about social science applications of Bayesian models including regression models, topic models, and social network models with an emphasis on foundational models. The course will build everything up from the basic principles, only requiring the knowledge of basic probability, statistics, and regression modeling along with the familiarity with R programming. The ultimate goal of this course is to teach fundamentals of Bayesian statistics that allow students to understand, implement, and even develop cutting-edge Bayesian models on their own.

1 Contact Information

Instructor

	Kosuke Imai
Office:	CGIS K306
Email:	imai@harvard.edu
URL:	https://imai.fas.harvard.edu/
Office Hours:	Mondays 1:30pm – 3:00pm (sign up at http://bit.ly/ImaiOfficeHours or reach out to me via Slack for an appointment)

Teaching Fellow

	Sooahn Shin
Email:	sooahnshin@g.harvard.edu
URL:	http://sooahnshin.com
Office Hours:	Wednesdays 4:00 – 6:00pm

2 Logistics

- Class meetings: Mondays, 9:45am–11:45am, CGIS K354 (except for Mar 20 at CGIS K031)
- Section meetings: Thursdays, 5:00pm – 6:15pm, TBD

3 Prerequisites

Students should have taken Gov 2001 and Gov 2002 or the equivalent courses in basic probability, statistics, data analysis, and regression modeling (e.g., Stat 110, Stat 111, and Stat 139).

4 Questions, Announcements, and Submissions

- The Canvas site for this course is at https://canvas.harvard.edu/courses/117655.

- We will use Gradescope at https://www.gradescope.com/courses/477983 for the submission of all assignments including the review questions and class exercises. There is a student guide you can check for any questions about the workflow.
- Rather than email, please use Ed Discussion at https://edstem.org/us/courses/31608 (the link is also available at Canvas) when asking questions about lectures, problem sets, and other course materials. This allows all students to benefit from the discussion and to help each other understand the materials. Both students and instructors are encouraged to participate in discussions and answer any questions that are posted. You may find this user guide helpful to orient yourself to the platform.
- Please feel free to use the Slack workspace for this course at https://gov-2017-s23-hpj.slack.com for any other communication with instructors and other students. You can join the Slack workspace from Canvas.
- A Google calendar that contains the information about the course logistics is available at this link.

5 Class Requirements

For each requirement, no late submission is allowed without a prior approval of the instructors.

- **Review questions (20% of final grade)** For the first three modules, you will work on one review question based on the reading assignment. You will be required to submit your answers by the beginning of the class meeting of the corresponding week. We will use the two best scores for each module to compute the final grade for the review questions. Although this means that you do not need to work on review questions for every week, we encourage you to try them in order to keep up with the course materials.
- **Exercises (30% of final grade)** For the first three modules, you will work on a couple of exercise questions. You will begin working on these exercise questions after the lecture during each class meeting. You will be required to submit your completed answers by the beginning of the section meeting of the corresponding week. We will use the two best scores for each module to compute the final grade for the exercises. Although this means that you do not need to work on exercises for every week, we encourage you to try them in order to keep up with the course materials.
- **Collaboration policy**: You may collaborate with other classmates and receive help from course staff on review questions and class exercises. **However, you must not copy anybody else's code or answers and are required to submit your own answer. Please specify the names of your collaborators at the beginning of your answer.**
- **Final project (40% of final grade):** The final project will be completed in collaboration with another student in the class. All projects must use a Bayesian methodology. Ideal projects will either (a) apply an existing technique to answer a substantive question or (b) extend such techniques in useful ways. Students are encouraged to consult the instructor and TF throughout the semester. Students should demonstrate that they can understand and implement a new Bayesian model and apply it to a data set of interest. Please note that we are not expecting you to fully execute an original research project! Rather, think of this as an opportunity to learn and apply a cutting-edge Bayesian model to a dataset of your choice. To help keep you on track, there will be multiple deliverables throughout the semester.
  - **February 20 (Project and collaborator identification)** By this date, pairs should submit a one-page project proposal with a brief statement of the problem to be solved or the question to be answered. Before the spring break, each pair should meet with the instructor to discuss the basic plan for the project and the methodological paper for their April presentation (see below).
  - **March 20 (First deliverable)** By this date, pairs must submit a first deliverable (maximum of 3 pages) including a concise problem statement, the short presentation of a descriptive analysis of data, and a brief explanation of possible Bayesian modeling strategy to be used. Students are

encouraged to meet with the instructor to discuss the direction of their project before their class presentation.

  - **April (Class presentation)** Each pair (or a group of pairs) will present a cutting-edge Bayesian model that forms the basis of their analysis. The presentation should consist of the detailed explanation of the model (15 minutes) and the results of their own implementation (5 minutes). The presentation will be followed by short Q&A. Everyone should have read the papers so that they also learn about the new Bayesian model and participate in the discussion.
  - **April 26 (Preliminary result)** By this date (the final day of classes), pairs must submit a PDF slide-deck (maximum of 10 slides) with the preliminary results of your analysis. Each student will be asked to comment on the slide-deck of another student.
  - **May 12 (Final project report)** By this date (last day of the exam period), pairs must submit the final report (no longer than 15 doublespaced pages) focusing on methods and results.

  Although these are fixed milestones, we are happy to supervise your project throughout the semester and beyond!

- **Participation (10% of final grade):** We will assess your overall engagement with the course materials throughout the semester (class and section meetings, online discussions).

6 Textbook

The textbook for the course is Gelman et al. (2013) *Bayesian Data Analysis* (3rd. Edition). Free electronic version available at http://www.stat.columbia.edu/~gelman/book/

7 Class Schedule

The course consists of four modules. In the first three modules, students will learn the foundation of Bayesian statistics through reading assignments, review questions, lectures, and class exercises. In the final module, students will use this foundation to learn new Bayesian models on their own. Through the final project, students will learn how to understand and implement new Bayesian models for empirical analysis.

Modules	Dates	Week	Topic	Readings	Review questions	Project milestone
1. Bayesian inference	1/23	0	Bayes' rule	BDA 1		
	1/30	1	Single parameter models	BDA 2	2.5	
	2/6	2	Multi-parameter models	BDA 3, 4	3.1	
	2/13	3	Hierarchical models	BDA 5, 6	5.7	Proposal due on 2/20
2. Bayesian computation	2/27	4	Posterior simulation I	BDA 10, 11	10.5	
	3/6	5	Posterior simulation II	BDA 12	12.4	
	3/20	6	EM Algorithm	BDA 13	13.9, 13.10	First deliverable due on 3/20
3. Bayesian regression	3/27	7	Regression models	BDA 14	14.1	
	4/3	8	Hierarchical regressions	BDA 15	15.6	
4. Bayesian modeling	4/10	9	Student presentations	TBD		
	4/17	10	Student presentations	TBD		
	4/24	11	Student presentations	TBD		Preliminary results due on 4/26

Some Potential Topics and Papers for the Final Projects

Below we list some potential topics and papers for the final projects and student presentation. Of course, students may choose other topics and papers in consultation with the instructor.

- Ideal point estimation

  Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. supreme court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002.

  Sean M. Gerrish and David M. Blei. How they vote: Issue-adjusted models of legislative behavior. *Neural Information Processing Systems*, 2012.

  Benjamin E. Lauderdale and Tom S. Clark. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771, 2014.

  Adam Bonica. Mapping the ideological marketplace. *American Journal of Political Science*, 58(2): 367–387, 2014.

  Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, Winter 2015.

  Kosuke Imai, James Lo, and Jonathan Olmsted. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656, December 2016.

- Topic models

  Jonathan B. Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, July 2008.

  Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256, 2009.

  Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

  Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

  Eshima Shusei, Kosuke Imai, and Tomoya Sasaki. Keyword assisted topic models. *arXiv 2004.05964*, 2022.

- Network models

  Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1097, December 2002.

  Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

  Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

  Kathryn Turnbull, Simón Lunagómez, Christopher Nemeth, and Edoardo Airoldi. Latent space modelling of hypergraph data. *arXiv preprint arXiv:1909.00472*, 2019.

  In Song Kim and Dmitriy Kunisky. Mapping political communities: A statistical analysis of lobbying networks in legislative politics. *Political Analysis*, 29(3):317–336, 2021.

Santiago Olivella, Tyler Pratt, and Kosuke Imai. Dynamic stochastic blockmodel regression for social networks: Application to international conflicts. *Journal of the American Statistical Association*, 117(539):1068–1081, 2022.

- Causal inference

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Corwin Matthew Zigler and Francesca Dominici. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107, 2014.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3): 965–1056, 2020.

Xun Pang, Licheng Liu, and Yiqing Xu. A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288, 2022.

- Nonparametrics

David B Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 2010.

Jacob M Montgomery, Santiago Olivella, Joshua D Potter, and Brian F Crisp. An informed forensics approach to detecting vote irregularities. *Political Analysis*, 23(4):488–505, 2015.

Daniele Durante, David B Dunson, and Joshua T Vogelstein. Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520):1516–1530, 2017.

JBrandon Duck-Mayr, Roman Garnett, and Jacob Montgomery. Gpirt: A gaussian process model for item response theory. In *Conference on Uncertainty in Artificial Intelligence*, pages 520–529. PMLR, 2020.

Eli Ben-Michael, David Arbour, Avi Feller, Alex Franks, and Steven Raphael. Estimating the effects of a california gun control program with multitask gaussian processes. *Annals of Applied Statistics*, forthcoming.