HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST 249 / BIOSTAT 249
Bayesian Methodology in Biostatistics
Spring 2020
Tuesdays & Thursdays, 3:45 – 5:15 pm, Kresge 202B

Instructor Information

Faculty
Jeffrey W. Miller
Assistant Professor of Biostatistics

Teaching Assistants
Alex Ocampo, 5th year PhD student, Biostatistics department

Xiao Wu, 3rd year PhD student, Biostatistics department

Credits
5 credits

Course Description
This course will cover the essential models, inference techniques, and basic theory of Bayesian analysis. We will cover classic models such as mixtures, GLMs, and HMMs, as well as admixtures, Dirichlet processes, and Gaussian processes. In addition to standard MCMC techniques, we will look at slice sampling, Hamiltonian Monte Carlo, and variational inference. Recent techniques for scaling to big data will also be covered. Case studies will be used to connect the concepts to real applications in the literature.

Pre-requisites
- BST 231 / BIOSTAT 231 (Statistical Inference I)
- BST 232 / BIOSTAT 232 (Methods I)
- Comfort with algorithms at the level of BIOSTAT 234 (Data Structures and Algorithms)

Outline of Topics / Learning Objectives

Upon successful completion of this course, students should understand the following topics:

- Introduction
  - Applications of Bayesian statistics, What is Bayesian statistics, Bayesian versus Frequentist perspectives
- Foundations
  - Prior and posterior, Marginal likelihood, Posterior predictive, Beta-Bernoulli example

- Conjugate priors
  - Exponential families, Conjugate families, Closed-form posterior calculations
- Gaussian models
  - Conjugate priors, Closed-form posterior calculations, Bayesian linear regression
- Monte Carlo approximation
  - Basic Monte Carlo, Importance sampling (IS), IS with unknown normalization constants
- Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling, Metropolis-Hastings, Markov chains, Combining MCMC moves, MCMC diagnostics, Slice sampling
- Mixture models (finite and infinite)
  - Finite mixtures, MCMC for mixtures, Dirichlet process mixtures (DPMs), MCMC for DPMs
- Admixture models
  - Latent Dirichlet Allocation (LDA), Population structure, MCMC for admixtures
- Variational inference (VI)
  - Classic VI using factorized approximations, VI for LDA, Automatic differentiation VI (ADVI)
- Model selection and Variable selection
- Hidden Markov models
  - Viterbi algorithm, Forward-backward algorithm, Baum-Welch algorithm
- Gaussian processes (GPs)
  - Examples of GPs, Positive semi-definite kernels (Covariance functions), GP regression, Inference in GPs
- Hamiltonian Monte Carlo (HMC) and Stan
  - HMC algorithm, No U-turn sampler (NUTS), Stan language, Illustration on hierarchical GLM
- Scaling to big data
  - Consensus Monte Carlo, Stochastic Variational Inference, GPs for big data
- Model building and model criticism
  - The art of likelihood construction, Prior selection, Model criticism, Robustness to misspecification

Course Readings

The lectures will primarily be based on my lecture notes, which will be available on the course website. Supplementary material is provided in the following textbooks:

- *Bayesian Data Analysis (Third Edition)*, by Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin. Chapman & Hall/CRC Press, 2013. **(REQUIRED)**
- *Pattern Recognition and Machine Learning*, by Christopher Bishop. Springer Science+Business Media, 2006. **(RECOMMENDED)**

Harvard Coop link to order books: https://tinyurl.com/300-W20-BIO-STAT-249-1

Course Website

Canvas Course Website: https://canvas.harvard.edu/courses/70601

Grading, Progress, and Assessment

The final grade for this course will be based on:

- Homework (50%)
- Case study presentation (10%)
- Class project – Midterm report (15%)
- Class project – Final report & presentation (25%)

Homework (50%)

Homework assignments will consist of problem sets and computer programming assignments. Assignments will be posted on the course website, and homework submission is mandatory. Homework will be graded by around one week after the due date.

Homework must be submitted online on the course website. Any handwritten material must be legible, otherwise no credit will be given. Kresge LL-19 and Countway Library have scanners that can be used at no cost. For programming exercises, you may use either R or Python, whichever you prefer. Include (a) plots and numerical results when appropriate, (b) discussion of the results when appropriate, (c) any supporting derivations, written out separately from the code, and (d) your source code (typed). The TAs will not run your code (e.g., to generate plots, etc.), so anything you want them to see must be included.

Policy on homework collaboration:

- Each student is required to come up with their own solutions for the homework.
- Students are allowed to discuss the problems in general terms (without sharing complete solutions) among themselves, or with the TAs or instructor. HOWEVER, when writing up their solutions, students are required to do this on their own, without copying from any other source.
- Students are forbidden from using solutions from any other source (such as solutions found online).
- Violation of this policy will result in a score of zero for that assignment, and possible disciplinary action.

Late submission policy:

- Homework submissions will be timestamped, and late submissions will be penalized as follows: starting from the due time until 48 hours after the due time, a multiplicative factor starting at 1.0 and decreasing linearly to 0.0 will be applied to the score. So, for example, the score for an assignment submitted 12 hours late will be multiplied by 0.75 (75% credit), 24 hours late will be multiplied by 0.5 (50% credit), and 48 hours late or later will be multiplied by 0.0 (no credit).
- There will be no make-ups or extensions.

Case study presentation (10%)

Groups of size 2-3 will take turns presenting case studies from the literature. Each group will present one time during the semester, and presentations will be scheduled periodically throughout the semester. Case study presentations will be around 25 minutes long and will occur during the regular lecture period.

For their presentation, each group will select a research article on an application of Bayesian statistics. A list of recommended articles will be provided, and any article not from this list must be pre-approved by the instructor. Presentations will be graded on clarity, correctness, completeness, and interestingness/fun. Further instructions will be provided on what to include in the presentations.

Normally, all group members will receive the same grade. If you feel that this would be unfair due to group members contributing unequally, the instructor must be informed in advance of the presentation.

Class project (Midterm 15%, Final 25%)

In groups of size 2-3 (not necessarily the same as the case study groups), students will engage in a semester-long project on a topic of their choosing. Topics must be relevant to the course content and must be approved by the course instructor.

Deliverables will consist of:

- Project proposal
  - 1 paragraph description of the project topic. Must be approved by instructor.
- Midterm paper & presentation
  - By midterm, each group must have completed a study of existing literature, obtained real data for application, and developed an initial idea of a new approach.
  - Partially complete paper: Abstract, Introduction, Previous work / background, Data, and References.
  - Midterm presentation: One class period will be designated for groups to take turns giving brief presentations on their projects.
- Final paper & presentation
  - Complete paper: Abstract, Introduction, Previous work / background, Proposed method, Simulation results, Empirical results on real data, and References.
  - Final presentation: The last two class periods of the semester will be designated for groups to take turns given presentations.

If you want to change topics, you must submit a new project proposal and have it approved by the instructor. If you change topics between the midterm and final, you will still need to submit a complete final paper.

Normally, all group members will receive the same grade. If you feel that this would be unfair due to group members contributing unequally, the instructor must be informed in advance of the presentation.

Course Schedule & Assessment of Student Learning

LN = Dr. Miller's lecture notes
BDA = Bayesian Data Analysis book
PRML = Pattern Recognition and Machine Learning book

Date	Planned Topic	Notes and Supplementary reading
Jan 28	Welcome, Introduction	LN 1, LN A (Prob and Lin Alg basics)
Jan 30	Foundations	LN 2, BDA 1.1-2.2
Feb 4	Conjugate priors	LN 3, BDA 2.3-2.7
Feb 6	Gaussian models	LN 4, BDA 3 (except 3.4 and 3.7), PRML 2.3.6-2.3.7
Feb 11	Bayesian linear regression	LN 5, BDA 14.1-14.2, 14.7, PRML 3.3
Feb 13	Monte Carlo approximation	LN 6, BDA 10, PRML 11.1
Feb 18	Markov chain Monte Carlo (MCMC)	LN 7, BDA 11, PRML 11.2-11.4
Feb 20	Markov chain Monte Carlo (MCMC)	" "
Feb 25	Stan language	LN 8, BDA Appendix C
Feb 27	Finite mixture models	LN 9, BDA 22
Mar 3	Dirichlet process mixture models	LN 10, BDA 23
Mar 5	Admixture models – Latent Dirichlet allocation	LN 11, Blei et al 2003
Mar 10	Admixture models – Population structure	LN 12, Pritchard et al 2000
Mar 12	Variational inference – Classic VI	LN 13, PRML 10.1-10.3, BDA 13.7
Mar 24	Variational inference – Automatic VI	LN 14, Kucukelbir et al 2015
Mar 26	Midterm project presentations	–
Mar 31	Hidden Markov models (HMMs)	LN 15, PRML 13.1-13.2
Apr 2	Hidden Markov models (HMMs)	" "
Apr 7	Gaussian processes	LN 16, BDA 21, PRML 6

Apr 9	Gaussian processes	" "
Apr 14	Model selection and Variable selection	LN 17, PRML 3.4, 4.4, Hoff 9
Apr 16	Advanced MCMC techniques	LN 18, BDA 12.1-12.3
Apr 21	Hamiltonian Monte Carlo and No U-turn sampler	LN 18, BDA 12.4-12.5, PRML 11.5
Apr 23	Scaling to big data	LN 19, Scott et al 2013, Hoffman et al 2013, Hensman et al 2013
Apr 28	Scaling to big data	" "
Apr 30	Model building and model criticism	LN 20, BDA 6, BDA 7
May 5	Model building and model criticism	" "
May 7	Breakout group discussions on selected topics	(Article on selected topic)
May 12	Final project presentations	–
May 14	Final project presentations	–

Please note, session topics and activities may be subject to change during the course.