QTM 385 - Introduction to Bayesian Statistics

Spring 2021

Instructor: Kevin McAlister

Email: kevin.mcalister@emory.edu

Synchronous Session Times: MW 11:20 AM - 12:35 PM EST

Synchronous Sessions: https://emory.zoom.us/j/95246416756

Office Hours: TBA

Office Hour Sessions: https://emory.zoom.us/j/97594430265

Course Description: This course is designed to introduce students to Bayesian methods for data analysis. While this course is largely an applied course, it is intended to provide modeling and computational tools to its students so that they will be able to implement the methods presented and develop new models for analyzing original data. The course will consist of three parts:

1) A review of probability theory and an introduction to the basic tools needed for Bayesian statistics including Bayes' theorem, maximum likelihood, prior elicitation, conjugacy, decision theory, model comparison, and other topics that form the basis of the recipe for Bayesian data analysis.

2) An overview of modern computational approaches used to perform statistical inference for Bayesian models including Markov Chain Monte Carlo methods, Hamiltonian Monte Carlo, variational approximations, and other computational tools that are used for Bayesian inference.

3) An introduction to modern problems in Bayesian statistics and the usage of Bayesian methods in machine learning. Topics discussed will include regression, regularization, latent variable estimation, and other important statistical domains where Bayesian methods provide solutions to long-standing problems.

Textbook: There will be one required textbook for this course:

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian Data Analysis.

This book (which will be referred to as BDA) is freely available from the authors here. Over the course of the semester, we will also be using a number of other resources that will posted to the course's Canvas site.

The reading for each week of the course is strongly recommended and should be completed at some point. However, there is no need to rush to complete each week's reading prior to the class meetings. There will be a few times over the semester where readings will be referenced heavily in the course discussions and I will make it clear in those weeks that they should be completed prior to the next course meeting.

Beyond BDA, there are numerous other references on the topics covered in this course that are freely available online. In many cases, these sites/books do a better job of explaining the material than BDA. I highly recommend seeking out these materials to better your understanding of the topics in this course. If you find anything particularly good, please send them my way so that I can post them for all the students in the course!

Prerequisites: As stated in the course listing, students must have completed all required QTM core courses before taking this class. This course will heavily leverage tools and topics in calculus and linear algebra, so competence with these topics is required. There is also an assumption that students have extensive exposure to probability and statistical inference. Students must also have some exposure to programming in R, Python, or another numerical computing language. While coding expertise is not required (you only get better with practice and this class will provide a lot of good directed practice!), familiarity with basic programming concepts is needed. I will be working mostly in R, but all assignments and projects can be completed in any appropriate programming language.

Course Structure: This course will be made up of lectures/discussions, problem-solving sessions, short individual quizzes, group problem sets, and a final group project.

1. **Synchronous Lectures** - There will be no formal attendance grade, but attendance is expected and strongly encouraged. The textbook and suggested readings are great resources for learning the materials, but there is no substitute for the face-to-face interactions that happen in class. Roughly 2 out of every 3 meetings will be lectures.

2. **Problem-solving sessions** - There will be no formal attendance grade, but attendance is expected and strongly encouraged. Roughly 1 out of every 3 lectures will be problem solving sessions to work on applied problems related to the course materials. These sessions will be used to work on problems similar to those in the problem sets or to introduce specific computational tools. These sessions will typically see students broken out into smaller groups to work on short ungraded assignments.

3. **Individual Quizzes** - There will be 5 individual quizzes posted on the Canvas website over the course of the semester that **will account for 25% of your final grade**. Your first quiz is the welcome quiz and you will be given full credit for this quiz upon completion. These quizzes will be two or three questions that are related to the materials covered since the last quiz. Quizzes will typically contain shorter versions of questions that are asked on the problem sets. Each quiz will be posted for one week and must be completed before the posted deadline. Since these are individual assignments, I ask that you complete these **on your own**. Pursuant to the Emory honor code, any violations of this request will result in a 0 for the assignment and, potentially, further escalation to ECAS.

4. **Group problem sets** - There will be 4 problem sets over the course of the semester that **will account for 35% of your final grade**. Problem sets will include a mixture of derivations, computational exercises, and applied data problems. Problem sets will typically be due 1.5 - 2 weeks after they are posted. Problem sets are to be completed with your QTM 385 groups - only one problem set needs to be turned in per group. I expect that students will work together closely within their groups to complete these assignments.

   Markdown or a similar typesetting method should be used for problem set solutions submissions. In R, RMarkdown is the best option. A short introduction to RMarkdown can be found here. For Python, Rmarkdown can be used with the `reticulate` package from RStudio or an equivalent markdown style submission can be made with a `.ipynb` file using Jupyter Notebooks. All homework submissions should be a `.zip` archive that includes, at a minimum, a `.Rmd` or `.ipynb` file and a rendered PDF version of the markdown file. If any external libraries are used, be sure to include the package `.tar.gz.` file in your `.zip` archive. Any non-base packages should be explicitly specified in a code block at the top of your write-up. There will be a few times throughout the semester where we will explicitly use other packages (Stan, `MCMCpack` in R, etc.), but the majority of your assignments should be completed with unique code - you will learn much more coding it yourself than using a pre-built package.

   I recognize that using Markdown and LaTeX-style math typesetting presents a bit of a learning curve in the beginning, but I believe that this is a valuable data science communication skill that is in demand. I will post a bit of a RMarkdown template for problem sets that can be used as a guide, but learning Markdown is largely a trial-and-error process. I will do my best to provide assistance when you run into problems.

5. **Final group project** - Each problem set group will also be tasked with completing a final group project that **will account for the other 40% of your final grade**. This project should apply the tools covered in this course to a novel problem that is of interest to the group. This can be analysis of a meaningful data set, a comparison of methods, development of a new method (methodological or computational), etc. The only rule is that the topic must be approved by the instructor (i.e. me). Your group should be thinking about this project early and often. In order to create a number of meaningful checkpoints, the final project has 4 parts:

   (a) **Initial Project Discussion (5%):** Before March 12th, 2020, each group should meet with me outside of class at least once to discuss their initial project ideas. This meeting can be used to get assistance with ideas or to let me know your final project topic. The goal of this meeting (or meetings, if need be) is to get approval on your final project topic.

   (b) **One-page project proposal (5%):** Before April 2nd, 2020, each group should submit a one-page project proposal that outlines your final project. This proposal should include your main question, data being used, methods, etc. I will provide comments on each proposal. At this stage, there should be no projects that do not meet course standards.

   (c) **Final project report (25%):** Before May 6th, 2020 at 12:00 PM EST, each group should submit their final project report. **This deadline is as late as possible in the semester and cannot be further extended!** This report should be written like an academic paper and report the question/motivation, methods, and findings. Any references should be properly cited. This paper should be **no more than 12 pages double-spaced and in 12 point font** including figures and tables, but not including your bibliography. This page limit is strict for two reasons - 1) an important skill in academia/industry is providing succinct reports of research findings and 2) longer papers in this stage of development tend to be unfocused reports that over-promise and under-deliver. Along with the report, replication files for any computational work should be submitted. These files should allow a researcher that does not have your data to reproduce your main findings (figures, tables, important quantities, etc.). Replication files can be submitted as a `.zip` archive or they can be uploaded to a Github project page. Final papers will be graded based on the quality of the work rather than the novelty of the findings - a strong attempt with null results is preferred to a low-quality paper with ”surprising” results.

   (d) **Group assessment (5%):** Before May 6th, 2020 at 12:00 PM EST, each student will fill out a survey that rates their own work over the course of the semester and rates the effort of their group members. The goal of this survey is to understand your feelings about how much work you put into the course, how much you learned, and how much effort each member of your group put into group problem sets and the final project. I expect that all students will receive full-credit for this assignment.

Group Assignment and Expectations: I am a firm believer in groups that are diverse in experience, background, culture, interests, etc. For this reason, all groups will be randomly assigned by me. Each group will consist of 3 members (and potentially 1 or 2 groups of 2). Because this course is being offered remotely, I will also be trying my best to assign groups where all members are in similar time zones. In the first week of classes, I will post an introductory survey on the course Canvas site where I will ask some general questions about time zone, interests, and statistical and programming backgrounds. I will use this info to assign groups before the first problem set is posted in the second week of the course.

Group assignments are fixed and will only be changed under extreme circumstances. Group members are expected to participate in all group assignments and share work in an equitable manner. As a way to ensure that grades are commensurate with effort and participation in group assignments, each assignment should include a statement at the top attesting to which students worked on the assignment. Please be sure to indicate if any group members did not meet your group's workload expectations. Note that this should be decided among your group - there is no competition for high grades, so there is no incentive to lie. Cases where group members do not fairly contribute will be dealt with on an individual basis. Should the situation arise where a group member is consistently shirking work, please let me know as soon as possible and we will work to come up with an equitable solution. If all else fails, the offending student will be required to do the rest of the course's work alone.

Late Assignments: Over the course of the semester, each student has 5 ”late days” that can be used to turn in group problem sets or individual quizzes after the due date. These late days are no questions asked, no notification needed, no grade penalty days to give students a grace period for late work over the semester. Late days are rounded up to the nearest day (an assignment due at 11:59 PM turned in at 12:01 AM on the next day counts as one late day). Group problem sets that are turned in late count as late days for all students in the group.

Should the situation arise where all late days are used by a student, then any assignment turned in late without my explicit approval will receive 50% credit. I recognize that there is a lot going on in the world right now and that each of us has our own unique set of things to deal with. If you find that you are unable to complete the course assignments on time, please reach out to me so we can come up something that is both fair and amenable to your unique situation.

Note that late days cannot be used for the final project and related checkpoints or the welcome quiz. I have provided ample time for each group to meet these deadlines, so all late final project assignments or welcome quizzes will be granted 50% credit.

Final Grades: As stated above (and stated again for emphasis), your final grade is made up of three components: individual quizzes (25%), group problem sets (35%), and the final group project (40%). Final grades will be determined using the following (estimated) grade ranges:

- **A**: 93% – 100%
- **A-**: 89% – 92%
- **B+**: 84% – 88%
- **B**: 80% – 83%
- **B-**: 75% – 79%
- **C+ and below**: < 75%

As assignment scores are computed and I get more of a feel for the average grades on each assignment, the grading rubric may be updated. However, the grade scale will only be loosened from this initial rubric (e.g. if your final average is 93%, then you will get an A for the course regardless of rubric changes). Final grade percentages will be rounded up to the nearest integer. There is no curve for final course grades - if everyone in the class scores high enough to get an A, then everyone will get an A. Hence, there is no ”competition” for grades.

Grade Appeals: If you believe that your grade on any assignment is incorrect or unfair, you should submit your concerns, in writing, to me. The written appeal should fully summarize what you believe the problems are and why. Unless the appeal regards a simple addition error, please wait 48 hours before submitting a written appeal.

Teaching and Learning During the Pandemic: Due to the unusual nature of the semester, communication is important. I will try my very best to respond to emails and questions posted to Canvas within 48 hours. In order to facilitate efficient communication, I ask that you post questions related to material and administrative policy to Canvas and to check Canvas and the syllabus before asking a question. If you have a question of a personal nature then please email me. I will likely be slower on weekends and it is usually not a great idea to ask questions on a Friday night or right before something is due.

If your situation changes regarding health, housing, or in any other regard with respect to your ability to participate in the class, please contact the appropriate Emory student support organization first and then me as soon as feasible. It is easier for me to address your needs if I know about them as soon as they arise. This does not mean I can successfully respond to every request for consideration, but I emphasize that my goal is to treat you all equitably and do what I can to help you succeed in this course.

Tentative Course Schedule:

Monday, January 25, 2021

Topics: Introductory Comments; Syllabus Review; Prerequisite refresher

Assignments: Welcome Quiz Posted

Wednesday, January 27, 2021

Topics: Prerequisite refresher; Maximum Likelihood; Monte Carlo Simulation

Friday, January 29, 2021

Assignments: Welcome Quiz Due by 11:59 PM EST

Monday, February 1, 2021

Topics: Bayes' Theorem[1]; Single Parameter Models; Multi-parameter Models

Assignments: Groups Posted; Problem Set 1 Posted

Wednesday, February 3, 2021

Topics: Problem Solving Session - Extending Bayes' Theorem to Data, Prior Choice and Sensitivity, 3 approaches to computing posteriors (analytic, approximation, and Monte Carlo)

Monday, February 8, 2021

Topics: The Exponential Family with special emphasis on the normal distribution; conjugate priors (and why conjugacy is still important even in the age of infinite computing power)

Wednesday, February 10, 2021

Topics: The multivariate normal distribution and posteriors in many dimensions

Friday, February 12, 2021

Assignments: Problem Set 1 Due at 11:59 PM EST; Quiz 2 Posted

Monday, February 15, 2021

Topics: Problem Solving Session - How informative is my prior? Uninformative, weakly informative, and reference priors.

[1] It's the Reverend Thomas Bayes' theorem, so it's possessive, but it is often omitted when writing about it.

Wednesday, February 17, 2021

Topics: No Class Meetings - Rest Day

Friday, February 19, 2021

Assignments: Quiz 2 Due at 11:59 PM EST; Problem Set 2 posted

Monday, February 22, 2021

Topics: Model comparison and the marginal likelihood (a.k.a. the dreaded denominator); Posterior Predictive Distribution; Some discussion on cross validation

Wednesday, February 24, 2021

Topics: Introduction to Bayesian Computation; Overview of approximations of posterior distributions, the advantages of Monte Carlo approximations, and Markov chain Monte Carlo.

Monday, March 1, 2021

Topics: Problem Solving Session - Metropolis-Hastings and Gibbs Sampling

Wednesday, March 3, 2021

Topics: Problem Solving Session - Slice Sampling and Automatic Bayesian Sampling

Friday, March 5, 2021

Assignments: Problem Set 2 Due at 11:59 PM; Quiz 3 Posted

Monday, March 8, 2021

Topics: Bayesian Linear Regression and Bayesian Probit Regression

Wednesday, March 10, 2021

Topics: Problem Solving Session - DIY Bayesian Regression

Friday, March 12, 2021

Assignments: Quiz 3 Due at 11:59 PM EST

Final Project Checkpoint: Groups must have met with me and had a tentative project idea approved before 5:00 PM

Monday, March 15, 2021

Topics: Hamiltonian Monte Carlo and its implementation via Stan.

Wednesday, March 17, 2021

Topics: Problem Solving Session - An Introduction to Stan.

Friday, March 19, 2021

Assignments: Problem Set 3 Posted

Monday, March 22, 2021

Topics: Hierarchical Models and Hyperpriors

Wednesday, March 24, 2021

Topics: Problem Solving Session - Rats and Argon in your Basement: The classic applied intro to Bayesian hierarchical regression with STAN

Monday, March 29, 2021

Topics: MAP estimates for Bayesian models; Regularization for large-p regression models

Wednesday, March 31, 2021

Topics: Problem Solving Session - Prove it to yourself: ElasticNet is just MAP Bayesian Estimation

Friday, April 2, 2021

Assignments: Problem Set 3 due at 11:59 PM EST; Quiz 4 Posted

Final Project Checkpoint: Groups must have submitted a one-page project proposal by 5:00 PM EST.

Monday, April 5, 2021

Topics: Catch-up day

Wednesday, April 7, 2021

Topics: An introduction to scalable Bayesian inference through variational approximations and a cautionary tale

Friday, April 9, 2021

Assignments: Quiz 4 Due at 11:59 PM EST; Problem Set 4 Posted

Monday, April 12, 2021

Topics: Latent Variables/Missing Data; An intro to the famous Latent Dirichlet Allocation model

Wednesday, April 14, 2021

Topics: No Class - Rest Day

Monday, April 19, 2021

Topics: Unsupervised Learning; Some latent variable modeling approaches

Wednesday, April 21, 2021

Topics: Problem Solving Session - DIY Finite Mixture Models/Factor Analysis/Probabilistic PCA

Friday, April 23, 2021

Assignments: Problem Set 4 Due at 11:59 PM EST; Quiz 5 Posted

Monday, April 26, 2021

Topics: Discussion - Causal Inference and The Deconfounder (Wang and Blei, 2019)

Wednesday, April 28, 2021

Topics: Discussion - Causal Inference in Panel Data and the generalized synthetic control method

Friday, April 30, 2021

Assignments: Quiz 5 Due by 11:59 PM EST

Thursday, May 6, 2021

Final Project Checkpoint: Final project report must be submitted by 12:00 PM EST; Group assessment must be completed by 12:00 PM EST