Final Paper Guidelines

Proposal Due: April 3, 2024 11:59PM

Final Paper Due: May 20, 2024 11:59PM

Objectives

The main objectives of the project are: (1) Demonstrate your understanding of Bayesian inference (2) Develop your skills in Bayesian computing and modeling (3) Provide you an opportunity to go beyond methods covered in class (4) Hone your skills in simulation and/or data analysis (5) Refine your technical writing skills.

Final Paper Options

There are two options for a final paper:

1. Data Analysis: conduct an original Bayesian analysis of a dataset of your choosing. The analysis could be one that applies an innovative Bayesian method to a data set previously analyzed using frequentist methods. The modeling should be creative - ideally extensions of models we covered in class.

2. Methodology: develop an original Bayesian method for a particular data problem. Conduct a simulation study comparing this method with alternatives. Do not be intimidated by "original" - you don't have to invent the next Gaussian Process Regression. Combining existing methods in new ways is still original.

Paper Proposal

A one-page proposal should be submitted declaring your paper option. If option 1, describe the data set you propose to analyze and confirm you are able to access it. If option 2, describe the methodological problem/gap you seek to address and an overview of your proposed method.

Publicly Available Data Sources

- Harvard Dataverse: `https://dataverse.harvard.edu/`
- Google Data search: `https://datasetsearch.research.google.com/`
- PLOSone journal: `https://journals.plos.org/plosone/browse/medicine_and_health_sciences`; All articles have a "data availability" section where some provide the data used in the article

Paper Structure

For both options, your paper should have the sections below. Your paper should use the LaTeXprovided to you. Paper should be no longer than 12 pages.

1. Introduction: explain the problem you are addressing and its associated statistical complexities. At a high level, explain your contribution. Distinguish your contribution from existing work on this problem: what will you be doing that is new/interesting?

2. Data Description and Structure: for option 1, explain your data set (is it from a trial, a prospective cohort study, what is the unit of observation, sample size, etc). Introduce appropriate mathematical notation to represent key features of the data. For option 2, your methodology is designed for a class of data structures - you should explain that data structure again introducing clear, concise mathematical notation.

3. Methodology: describe the methodology/models that you are applying and/or developing in detail. Explain how it handles the statistical complexities mentioned in the introduction. Discuss any limitations. Discuss/justify prior distributions and hyperparameter choices.

4. Computation: detail a posterior sampling and computation strategies needed for your model. This should be some variation of a metropolis-in-gibbs sampler or something fancier, but the steps should be detailed. I.e. you can't just say "we used Stan to do it." You should write down explicitly the posterior distribution up to a proportionality constant (i.e. likelihood times prior).

5. Data Analysis (for Option 1 only): provide summary statistics of the data relevant to your analysis. Present your results (usually in the form of tables/visualizations, etc), results of posterior predictive/fit checks, and conclusions. Be sure to explicitly tie in your results to the problem you described in the Introduction. Be sure to detail choices of analysis - ranging from what variables you used to number of MCMC iterations, etc. Conduct appropriate MCMC convergence checks.

6. Simulation Study (for Option 2 only): describe the data generating process underlying each simulation setting. Describe all the parameters of the simulation study (how many iterations, what was the sample size of the data generated in each iteration, etc). What are the performance metrics you decided to evaluate (e.g. out of sample AUC, bias, variance, MSE, coverage, etc) and why did you choose those? What comparators methods did you choose and why? Present (usually in a table) and discuss the results and state your conclusions.

Evaluation Criteria

There will be 100 points available. Allocation and approximate evaluation criteria are as follows:

- Introduction (10 pts): problem and associated statistical complexities are clearly explained. The proposed work is clearly explained and contributions distinguished from previous work.
- Data Description and Structure (10 pts): data structure is and appropriate mathematical notation is introduced to represent the data. Notation introduced should be concise and unambiguous.
- Methodology (30 pts): the methodology developed/used is appropriate for the data structure of the problem and appropriately handles complexities discussed in the introduction. For option 1: models that go beyond ones we have covered in class are more likely to rank high in innovation. More standard/routine approaches are less likely to rank high in innovation.
- Computation (20 pts): the posterior distribution (up to proportionality constant) correctly written. The computational strategy valid and clearly described.
- Data Analysis / Simulation Study (30 pts): results are presented clearly and concisely with legible visualizations/tables. Results are interpreted appropriately. Conclusions are justified by the model and results. All analyses are reproducible with code provided (ideally in an accompanying GitHub repository that is linked in the paper).