INTRODUCTION TO BAYESIAN DATA ANALYSIS (STAT3016/4116/7016)

SEMESTER 2 2021

FINAL PROJECT

DUE DATE: Thursday 4 November 2021, by 11:59pm

(40% of total course grade)

GENERAL DESCRIPTION

For the final project you will analyse a dataset of your choice using any appropriate Bayesian method(s) we have discussed in class. You will need to formulate your own research question(s), and apply your knowledge of Bayesian statistics and computational strategies to answer your chosen question(s).

The dataset could be one of academic or personal interest to you, and that fits one of the Bayesian methods discussed in class (or some closely related Bayesian method that you can implement). Your chosen dataset should be a real data set which you have not analysed before and which has not been analysed in a textbook. Once you have chosen your real data set, it is recommended that you confirm with the lecturer that your choice of data set is suitable for the final project in terms of complexity (number of records and number of variables) and potential research questions which may be answered using Bayesian methods.

The following website may be useful for you in selecting a data set for your final project:

- UC Irvine Machine Learning Repository `https://archive.ics.uci.edu/ml/index.php`

Alternatively, a couple of datasets will be posted on the course website and you may choose to use one of these for your project. You do not need to seek prior approval if you use one of these data sets.

The final project is to be done individually. University policies on plagiarism will be **strictly** enforced.

You must submit a written report to communicate your project findings. Your report must be submitted electronically via Turnitin on the course website.

Please include the following sections in your report:

- **Introduction**: Explain the motivation behind your research question and state the dataset you are using (including its source). What are the main issues or problems to be addressed? Why is this research question of interest to you?
- **Methodology**: Describe the variables to be used in your analysis. Specify and justify your choice of prior distribution(s) and sampling model(s) and derive your posterior distribution(s). Be sure to define all notation and state any assumptions you make. If you are implementing an MCMC algorithm, please provide step by step details of your algorithm (that is, the sequence of draws at each iteration). Be sure to specify the number of iterations, burn-in period or thinning interval, and choice of proposal distributions (if applicable). Specify what software package(s) you used to implement your Bayesian model, and whether you wrote your own code or used an existing package.
- **Results**: Describe the main findings of your analysis. Be sure to relate your discussion on your results back to your original research question(s). Include graphs if appropriate and the results of any model checking and MCMC convergence diagnostics which you performed.
- **Conclusions**: Summarize the main findings of your project. What did you learn? What are the key points that a reader should take away? Discuss the strengths of your analysis. Discuss any limitations of your analysis, for example, did you need to make any simplifying assumptions when deriving your model or hack your code to get something working? Briefly describe any next steps that you would take to extend or improve your analysis if you had more time or additional resources.
- **Appendices**: Attach the main computer code files for your analysis. If applicable you may also include any detailed mathematical derivations in the appendices that do not need to be contained in the main body of the report.
- **Reference List**: If applicable. (see here for standard referencing styles `https://www.anu.edu.au/students/academic-skills/academic-integrity/referencing`

Total length: 15-20 pages (excluding appendices but including graphs).

PROJECT PROPOSAL

You may submit a short project proposal to the course lecturer no later than Friday 17 September 2021 (or earlier). The project proposal is not graded. It exists primarily for you to get feedback on your project idea and to make sure you have started thinking about your project. The proposal should comprise up to one page addressing the following questions:

- Which data set are you using?
- What are the main issues or problems to be addressed?
- What variables in the data set will you use?
- What are your initial thoughts on appropriate models/distributions?
- What questions and/or concerns do you have about the project?

PROJECT GRADING GUIDELINES

In addition to correct specification of your model and the depth of your analysis, the marker will also be looking for the following in your report:

1. **Consistency**: Did you answer your question(s) of interest?
2. **Clarity**: Is it easy for the reader to understand what you did and the arguments you made? Is the report logically structured?
3. **Relevancy**: Did you use Bayesian statistical techniques wisely to address your question?
4. **Interest**: Did you tackle a challenging, interesting question?
5. **Methodology**: Are all components of your model clearly described? Have you defined all your notation? Have you included step by step details of your MCMC algorithm? In other words, is it straightforward for a person to implement your model after reading your report?

Some tips:

- State your question(s) up front, and use statistical modelling to help answer it. The models should not drive the question; the question should drive the models.
- Talk to the teaching staff for advice.
- Be selective with computer output in the appendix to help clarity.
- If you are using techniques we learned in class, you do not need to re-explain the theory behind the techniques. If you are using techniques that we did not cover in class, the techniques should be clearly explained in your report.

Final Project grade breakdown (subject to minor changes):

Item	**Total marks available**
Consistency/Clarity/Graphical Displays	7
Relevancy	5
Interest	8
Methodology	10
Results + Discussion	10
Model Diagnostics	5
Conclusions	5
Total	**50**