Instructions – project work

Choose a data set and make all the steps of Bayesian data analysis workflow as explained below.

Project notebook

The Project outcome is a R or Python notebook similar to notebooks in

- BDA R demos https://github.com/avehtari/BDA_R_demos/tree/master/demos_rstan
- BDA Python demos https://github.com/avehtari/BDA_py_demos/tree/master/demos_pystan
- Stan case studies http://mc-stan.org/users/documentation/case-studies.html
- StanCon case studies http://mc-stan.org/users/documentation/case-studies.html (some of these notebooks are for a bigger projects, but reflect still the basic idea of a notebook presentation)

The submitted notebooks need to illustrate the knowledge of the Bayesian workflow. It has to include:

- Description of the data, and the analysis problem
- Description of at least two models, for example:
  - non-hierarchical and hierarchical
  - linear and non-linear
  - variable selection with many models
- Informative or weakly informative priors, and description of the prior choices
- Stan code
- How Stan model is run
- Convergence diagnostics (Rhat, divergences, ESS)
- Posterior predictive checking
- Model comparison (e.g. with loo)
- Predictive performance assessment if applicable (e.g. classification accuracy)
- Sensitivity analysis with respect to prior choices
- Discussion of problems, and potential improvements

Group composition:

- We generally recommend 2 person groups
- 1 or 3 person groups are also fine if not otherwise possible or sensible
- 3 person groups are expected to choose more difficult projects
- 2-3 person groups are highly recommended over 1 person groups. 2-3 person groups have priority when reserving presentation slots.

Project Presentation

Presentation details:

- Each project needs to be presented in addition to submitting the notebook
- The presentation should be high level but sufficiently detailed information should be readily available to facilitate answering questions from the audience
- Within each session, about four groups will be presenting
- For 1-2 person groups, the presentation should be 10 minutes
- For 3 person groups, the presentation should be 15 minutes
- Afterwards, questions will be asked first by other students and then by two attending TAs for about 5 to 10 minutes
- Grading of the presentation will be done by the two TAs using standardized grading instructions
- Presenters' ID cards will be checked to ensure the right persons are presenting

Specific presentation recommendations:

- The first slide needs to include project title and names of the group members.
- The chosen statistical model(s), including both likelihood and priors, should be explained and justified (you are *not* holding this presentation for a hypothetical customer who doesn't care about the details of your models).
- Big enough font size for text and figure labels should be used so that slides are easily readable for the audience.
- The last slide needs to include to conclusion and names of the group members.
- If the last slide has only "Thanks", "Thank you", "Questions?" or something similar of zerocontent, the project grade is reduced by one. So we recommend you avoid such slides (in general).
- The best presentations we have seen were by groups who did visit TA session(s) and showed intermediate results to TAs and did get feedback how to improve. So we recommend you visiting TA sessions for the project.

Data sets and modelling recommendations

Suggestions for sensible data sets:

- TODO

Forbidden Datesets: Some data sets should not be used in the project:

- Titanic (R data set)
- mtcars (R data set)
- baseball batting (the one used in Bob Carpenter's case study)

Modelling requirements and recommendations:

- Every parameter needs to have an explicit proper prior. Improper flat priors (default in Stan) are not allowed.
- An hierarchical model is a model, where the prior of certain parameters contains other parameters, which are also estimated in the model. For example. `b ~ normal(mu, sigma)` with `mu ~ normal(0, 1)` and `sigma ~ exponential(1)`
- If a parameter does not have both a natural lower and upper boundary, `uniform(a, b)` are considered bad practice as they set hard boundaries on parameters without such natural hard boundaries.
- Models should ideally use at least one covariate. That is, just modeling your outcome variable without doing any predictions or similar using other variables, is likely too simple of an analysis for this project.
- Using brms for the model specification is allowed, but the Stan code needs to be included, briefly commented, and all priors need to be checked from the Stan code.

Peergrade rubric

Part of the questions are used to check that the minimal requirements of the project work are included in the notebook. Most of the questions are for giving feedback to other students. The received feedback and your response to that will be discussed in the evaluation meeting.

- Can you open the notebook?
  - yes
  - no
- Is there an introduction?
  - There is no clear introduction
  - The introduction touches on the main topic
  - The introduction states the main topic and provides an overview of the notebook
  - The introduction is inviting, presents an overview of the notebook. Information is relevant and presented in a logical order.
- What are your suggestions on how to improve the introduction?
- Is there a conclusion?
  - There is no clear conclusion
  - A conclusion is included
  - The conclusion is clear

  Describe in your own words what is the main conclusion of the data analysis in this notebook?
- The structure and organization of the notebook

  - The notebook lacks a clear data analysis story
  - The notebook attempts to tell a coherent data analysis story but lacks some focus and clarity.
  - The notebook presents a clear cohesive data analysis story
  - The notebook presents a clear cohesive data analysis story, which is enjoyable to read
- Overall, what did you think of the structure and organization of the notebook? Name at least one way your peer could improve structure and organization.
- Accuracy of use of statistical terms
  - There are numerous errors in use statistical terms
  - There are some errors in use of statistical terms
  - Statistical terms are used accurately but sometimes lack clarity
  - Statistical terms are used accurately and with clarity
- Description of the data, and the analysis problem
  - yes
  - no
  - Did you get a sense of what is the data and the analysis problem when they were first introduced? Where and how might the author make the model description more clear?
- Are there more than one model
  - yes
  - no
  - Was it easy to find the list of the models?
- Description of the models
  - yes
  - no
  - Did you get a sense of what the models are? Where and how might the author make the model description more clear?
- Description of the prior choices
  - No priors or improper priors (e.g. uniform on unconstrained parameter) used
  - Priors listed but not justified
  - Priors are listed and justified
- Is Stan code included?
  - yes
  - no

- Is the code for how Stan model is run included?
  - yes
  - no
- Is Rhat convergence diagnostics included?
  - No
  - Yes, but no discussion what can be concluded from the shown Rhat values
  - Yes, with discussion what can be concluded from the shown Rhat values
- Are HMC specific convergence diagnostics (divergences, tree depth) included?
  - No
  - Yes, but no discussion what can be concluded from the shown values
  - Yes, with discussion what can be concluded from the shown values
- Is effective sample size diagnostic (ESS) included?
  - No
  - Yes, but no discussion what can be concluded from the shown values
  - Yes, with discussion what can be concluded from the shown values
- Is there posterior predictive checking?
  - No
  - Yes, but no discussion what can be concluded from the shown checks
  - Yes, with discussion what can be concluded from the shown checks
- Are there sensitivity analyses with respect to likelihood and/or prior choices?
  - No
  - Yes, but no discussion what can be concluded from the shown sensitivity analyses
  - Yes, with discussion what can be concluded from the shown sensitivity analyses
- Is there a discussion of problems and potential improvements ?
  - No
  - Some
  - Very clear
- Choose something you like about the notebook and explain why you like it.
- If you were to go back and redo your own notebook after reading this submission, what would you change?
- If the student(s) were to complete this project work again, what could they change, to make it overall better?