Project work details

Project work involves choosing a data set and performing a whole analysis according to all the parts of Bayesian workflow studied along the course.

- The project work is meant to be done in period II.
- In the beginning of the period II
  - Form a group. We prefer groups of 2-3, but the project can be done alone.
  - Select a topic. You may ask in the course chat channel #project for opinion whether it's a good topic and a good dataset. You can change the topic later.
  - Start planning.
- The main work for the project and the presentation will be done in the second half of the period II after all the workflow parts have been discussed in the course.
- The online presentations will be made on the evaluation week after period II.

Project schedule

- Form a group and pick a topic. Register the group before November 9 23:59.
- Groups of 2-3 can reserve a presentation slot starting November 11.
- Groups of 1 can reserve a presentation slot starting November 11.
- Groups that register late can reserve a presentation slot starting 14th November.
- Work on the project. TA session queue is also for project questions.
- Project report deadline December 6. Submit in peergrade (separate "class", the class code will be posted in the course chat announcements).
- Project report peer grading December 7-9 (so that you'll get feedback for the report before the presentations).
- Project presentations December 14-18.

Groups

Project work is done in groups of 1-3 persons. Preferred group size is 2-3, because you learn more when you talk about the project with someone else.

If you don't have a group, you can ask other students in the group chat channel **#project**. Tell what kind of data you are interested in (e.g. medicine, health, biological, engineering, political, business), whether you prefer R or Python, and whether you have already more concrete idea for the topic.

Groups of 2-3 students can choose their presentation time slot before 1 student groups. 3 person group is expected to do a bit more work than 2 person group.

You can do the project alone, but the amount of work is expected to the same for 2 person groups.

TA sessions

The groups will get help for the project work in TA sessions. When there are no weekly assignments, the TA sessions are still organized for helping in the project work.

Evaluation

The project work's evaluation consists of

- peergraded project report (40%) (within peergrade submission 80% and feedack 20%)
- presentation and oral exam graded by the course staff (60%)

- clarity of slides + use of figures
- clarity of oral presentation + flow of the presentation
- all required parts included (not necessarily all in main slides, but it needs to be clear that all required steps were performed)
- accuracy of use of terms (oral exam)
- responses to questions (oral exam)

Project report

In the project report you practice presenting the problem and data analysis results, which means that minimal listing of code and figures is not a good report. There are different levels for how data analysis project could be reported. This report should be more than a summary of results without workflow steps. While describing the steps and decisions made during the workflow, to keep the report readable some of the diagnostic outputs and code can be put in the appendix. If you are uncertain you can ask TAs in TA sessions whether you are on a good level of amount of details.

The report should include

1. Introduction describing
   - the motivation
   - the problem
   - and the main modeling idea.
   - Showing some illustrative figure is recommended.
2. Description of the data and the analysis problem. Provide information where the data was obtained, and if it has been previously used in some online case study and how your analysis differs from the existing analyses.
3. Description of at least two models, for example:
   - non hierarchical and hierarchical,
   - linear and non linear,
   - variable selection with many models.
4. Informative or weakly informative priors, and justification of their choices.
5. Stan code (brms can be used to generate the code, but Stan code needs to be present and explained).
6. How to the Stan model was run, that is, what options were used. This is also more clear as combination of textual explanation and the actual code line.
7. Convergence diagnostics ($\widehat{R}$, ESS, divergences) and what was done if the convergence was not good with the first try.
8. Posterior predictive checks and what was done to improve the model.
9. Model comparison (e.g. with LOO-CV).
10. Predictive performance assessment if applicable (e.g. classification accuracy) and evaluation of practical usefulness of the accuracy.
11. Sensitivity analysis with respect to prior choices (i.e. checking whether the result changes a lot if prior is changed)
12. Discussion of issues and potential improvements.
13. Conclusion what was learned from the data analysis.
14. Self-reflection of what the group learned while making the project.

Project presentation

In addition to the submitted report, each project must be presented by the authoring group, according to the following guidelines:

- The presentation should be high level but sufficiently detailed information should be readily available to help answering questions from the audience.

- The duration of the presentation should be 10 minutes (groups of 1-2 students) or 15 minutes (groups of 3 students).
- At the end of the presentation there will be an extra 5-10 minutes of questions by anyone in the audience or two members of the course staff who are presenr. The questions from lecturer/TAs can be considered as an oral exam questions, and if answers to these questions reveal weak knowledge of the methods and workflow steps which should be part of the project, that can reduce the grade.
- Grading will be done by the two members of the course staff using standardized grading instructions.

Specific recommendations for the presentations include:

- The first slide should include project's title and group members' names.
- The chosen statistical model(s), including observation model and priors, must be explained and justified,
- Make sure the font is big enough to be easily readable by the audience. This includes figure captions, legends and axis information,
- The last slide should be a summary or take-home-messages and include contact information or link to a further information. (The grade will be reduced by one if the last slide has only something like "Thank you" or "Questions?"),
- In general, the best presentations are often given by teams that have frequently attended TA sessions and gotten feedback, so we strongly recommend attending these sessions.

More details on the presentation sessions

- If you don't have microphone or video camera (e.g. in your laptop or mobile phone) then we'll arrange your presentation on campus in period III.
- If you reserved a presentation slot but need to cancel, do it asap.
- Zoom meeting link for all time slots available in the course chat.
- As we have many presentation in each slot join the meeting in time. Late arrivals will lower the grade. Very late arrivals will fail the presentation and can present in period III.
- Presenting group needs to have video and audio on.
- It is easiest if just one from the group shares the slides, but it is expected that all group members present some part of the presentation orally.
- Presentation time is 10 min for 1-2 person groups and 15min for 3 person groups
- Time limit is strict. It's good idea to practice the talk so that you get the timing right. Staff will announce 2min and 1min left and time ended. Going overtime reduces the grade.
- After the presentation there will be 5min for questions, answers, and feedback.
- Each student has to come up with at least one question during the session. Students can ask more questions. Questions by students are posted in chat, and they can be posted already during the presentation.
- Staff Will ask further questions (kind of oral exam)
- Grading of the project presentation takes int account
  - clarity of slides + use of figures
  - clarity of oral presentation + flow of the presentation
  - all required parts included (not necessarily all in main slides, but it needs to be clear that all required steps were performed)
  - accuracy of use of terms (oral exam)
  - responses to questions (oral exam)
- Students will also self-evaluate their project. After the presentation each student who just presented sends a private message to one of the staff members with a self evaluation grade from themselves and for each group member (if applicable).

Data sets

As some data sets have been overused for these particular goals, note that the following ones are forbidden in this work (more can be added to this list so make sure to check it regularly):

- R data sets titanic and mtcars

- Baseball batting (used by Bob Carpenter's StanCon case study).
- Data sets used in the course demos

It's best to use a dataset for which there is no ready made analysis in internet, but if you choose a dataset used already in some online case study, provide the link to previous studies and report how your analysis differs from those (for example if someone has made non-Bayesian analysis and you do the full Bayesian analysis).

Depending on the model and the structure of the data, a good data set would have more than 100 observations but less than 1 million. If you know an interesting big data set, you can use a smaller subset of the data to keep the computation times feasible. It would be good that the data has some structure, so that it is sensible to use multilevel/hierarchical models.

Model requirements

- Every parameter needs to have an explicit proper prior. Improper flat priors are not allowed.
- A hierarchical model is a model where the prior of certain parameter contain other parameters that are also estimated in the model. For instance, `b ~ normal(mu, sigma)`, `mu ~ normal(0, 1)`, `sigma ~ exponential(1)`.
- Do not impose hard constrains on a parameter unless they are natural to them. `uniform(a, b)` should not be used unless the boundaries are really logical boundaries and values beyond the boundaries are completely impossible.
- At least some models should include covariates. Modelling the outcome without predictors is likely too simple for the project.
- `brms` can be used, but the Stan code must be included, briefly commented, and all priors need to be checked from the Stan code and adjusted to be weakly informative based on some justified explanation.

Some examples

The following case study examples demonstrate how text, equations, figures, and code, and inference results can be included in one report. These examples don't necessarily have all the workflow steps required in your report, but different steps are illustrated in different case studies and you can get good ideas for your report just by browsing through them.

- BDA R and Python demos are quite minimal in description of the data and discussion of the results, but show many diagnostics and basic plots.
- Some Stan case studies focus on some specific methods, but there are many case studies that are excellent examples for this course. They don't include all the steps required in this course, but are good examples of writing. Some of them are longer or use more advanced models than required in this course.
  - Bayesian workflow for disease transmission modeling in Stan
  - Model-based Inference for Causal Effects in Completely Randomized Experments
  - Tagging Basketball Events with HMM in Stan
  - Model building and expansion for golf putting
  - A Dyadic Item Response Theory Model
  - Predator-Prey Population Dynamics: the Lotka-Volterra model in Stan
- Some StanCon case studies (scroll down) can also provide good project ideas.