Handelgroup Bayesian Modeling Protocol

Contributors: Zane Billings, Andreas Handel.

Last updated: 2024-08-05

This document provides an overview of the general protocol for a computational project in Handelgroup which will implement a Bayesian modeling computational analysis.

Materials and Equipment

- A workstation, preferably with >4 cores and 16GB RAM.
- One of the following software setups:
  - R version 4.1 or newer, along with either cmdstan version 34 or newer or the current CRAN version of rstan.
  - An appropriate version of Julia with Turing.jl or Python with PyMC5.
  - Another appropriate software package for Bayesian modeling approved by the research team.
- A working Quarto installation.

Safety

There is no expected risk to researchers other than risks associated with a typical non-strenuous office job. Prolonged sitting and/or viewing a computer screen can have long-term detrimental health consequences. Research personnel should have ergonomic desk setups and frequent stretching, movement, and altering position is recommended. No PPE is required.

Methods

1. Project organization: all projects should begin in an R project based on the Handelgroup Data Analysis Project Template.

   a. Optionally, renv can be used for package management. Regardless of how package management is handled, the final work product should contain a description of which packages and versions were used to run the code.
   b. The repository should be linked to Git/GitHub and updates should be committed and pushed to GitHub frequently. All project repositories for papers should be stored in the `ahgroup` GitHub organization.
2. Outcome identification: the modeling outcomes and hypotheses should be clearly explained in the introduction and methods sections of the text, and should be identified before modeling begins.

   a. Updates may occur in an iterative process throughout the model development process.
   b. While not compulsory, a Directed Acyclic Graph (DAG) is recommended for the identification of control variables in modeling.

   c. The likelihood function for the outcome and the functional structure of the parameters should be informed by underlying science or good statistical practice.
3. Data cleaning: all data cleaning should be conducted in R. The final data file used for analysis should be saved in serialized R-data (i.e., Rds) format and clearly labeled.
4. Exploratory data analysis: all EDA should be conducted in R and documented in either an R script or Quarto file, unless the research time agrees on the use of another software.

   a. Appropriate descriptive statistics, such as counts, percentages, measures of central tendency, measures of spread and dispersion, etc. should be calculated in order to better understand the data.
   b. In addition to the calculation of descriptive statistics, data should be visualized during the EDA process.
5. Formal Bayesian modeling: Bayesian modeling should primarily be done using Stan (including R packages which interface with Stan). Another Bayesian PPL can be used instead if the research team agrees.

   a. The likelihood and functional form of any parameter models for the outcome should be based on underlying biology first. Good statistical practice can be used to inform model development as well (e.g., using a Student's t likelihood to model a noisy outcome with outliers).
   b. Functional and hierarchical forms of covariates should be well-documented in either the methods or the supplementary material of any papers. This can take the form of `lme4`-style regression syntax, `brms` extended syntax, or a statistical model.
   c. Priors should be chosen based on good statistical practice. We recommend the use of prior predictive simulation to determine sensible priors. The LKJ Cholesky parametrization of covariance/correlation matrices should always be used if available.
6. Result presentation: the posterior predictions and derived outcomes should be explained in enough detail that a reader can understand which predictions were generated from a Bayesian model (e.g. link or response scale), any transformations done to those predictions, and any calculations used to derive outcomes like causal estimands. For hierarchical models, the supplement should clarify which methods were used to obtain marginal or conditional credible intervals for presented outcomes.