

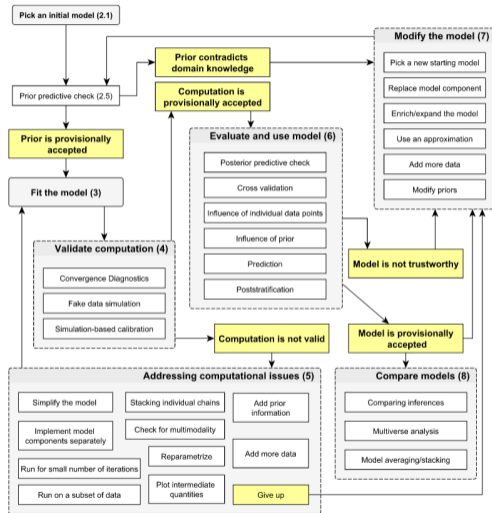
CS-E407520 - Special Course in Machine Learning and Data Science: Bayesian Workflows

Session 1: Course practicalities & Choosing an initial model

April 22, 2024



Welcome to the course!



Bayesian workflow by Gelman et al. (2020)

Schedule for today's session

Time	Activity
20 min	Introduction to the course
20 min	Discussion of modelling problems & datasets
10 min	Break
15 min	Primer for next workflow steps
30 min	Getting started

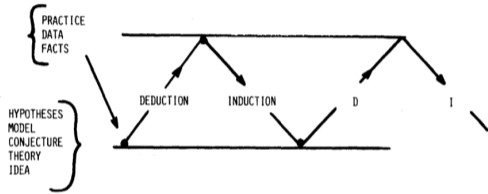
Introduction to the course

Motivation: Modelling as learning

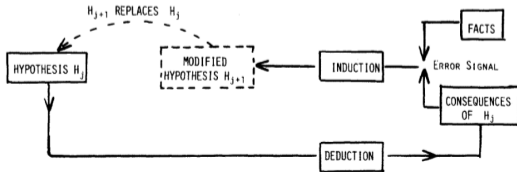
A. The Advancement of Learning

A(1) An Iteration Between Theory and Practice

A(2) A Feedback Loop



- Modelling as a device for learning by connecting theory and observations
- Model development is iterative



Box (1976)

Motivation: Modelling as software development

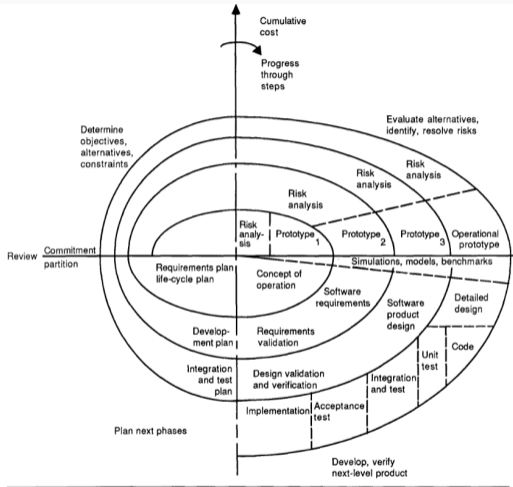


Figure 2. Spiral model of the software process.

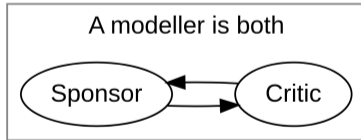
- Model building involves a range of decisions and tasks
- Modelling often starts with a prototype that is refined
- Modelling and computation are intertwined

Boehm (1988)

Motivation: Why Bayesian workflows?

Modelling workflows can help with

- Guidance: navigating necessary and optional decisions and tasks
- Knowledge transfer: explicit recommendations and accessible expertise
- Transparency: reproducible modelling procedures and transparent modelling choices



What to expect: Goals of the Seminar

You will ...

... learn about different **tasks in Bayesian workflows**

... gain **practical experience** with Bayesian workflow steps

... get familiar with **software tools** for supporting modelling workflows

... actively **apply what you learn** to your data analysis project

What to expect: Structure of the Course

Sessions	Content
Session 1	Course practicalities & primer “Choosing an initial model”
Session 2	Workflow diaries & primer “Prior Choices”
Session 3	Workflow diaries & primer “Model checking”
Session 4	Workflow diaries & primer “Extending Models & Model Selection”
Session 5	Workflow diaries & primer “Interpreting and Presenting Model Results”
Session 6	Workflow diaries & Q&A for preparing presentations
Session 7	Project presentations & summary of the course

What to expect: **Approach to Teaching**

- Strong focus on **discussion of your projects** and independent work and learning
- Seminar sessions will not have lecture/tutorial focus
- Students will also learn and apply concepts outside of sessions

What to expect: Independent Learning

- You will be given a range of resources to work through and apply to your analysis problem during the week
- Some support for small queries through Zulip ([click here to join](#))
- Extended discussion/support will then take place in the following seminar session

What to expect: Interactive Seminars

Time	Activity
45 min	Discussion of workflow diaries
10 min	Break
35 min	Discussion of workflow diaries (ctd.)
15 min	Primer for next workflow steps

What to expect: Interactive Seminars

- Weekly sessions will primarily be interactive presentations and discussion
 - Each student will share their experiences applying the workflow step during the previous week
 - Staff will make requests/suggestions for students to investigate during presentation
 - Other students can see different ways to apply workflow step to different problems
- Session will conclude with small 'primer' introduction to concepts of next workflow step
- Relevant resources for the week then provided

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course
 - Used during each session to share status of project and apply suggestions/requests

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course
 - Used during each session to share status of project and apply suggestions/requests
 - Presented in Session 7

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course
 - Used during each session to share status of project and apply suggestions/requests
 - Presented in Session 7
 - Can be used as a reference for future analyses

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course
 - Used during each session to share status of project and apply suggestions/requests
 - Presented in Session 7
 - Can be used as a reference for future analyses
- You need to have a **dataset and research question** to walk through Bayesian workflow steps during the course

Assessment: Workflow Diary

- You will maintain an **interactive record of the code and results** from applying the workflow steps throughout the course
 - Used during each session to share status of project and apply suggestions/requests
 - Presented in Session 7
 - Can be used as a reference for future analyses
- You need to have a **dataset and research question** to walk through Bayesian workflow steps during the course
- If you do not have a dataset yet, several options are listed on the course website:
Datasets

Assessment: Workflow Diary

Two templates are provided for the diary

- Quarto ([link](#))
- Jupyter Notebook ([link](#))

i Note

You can also choose to maintain the record in a form of your choice as long as this allows you to interactively update the results/code during presentations in seminar sessions.

Assessment: Discussions in the Seminars

- Active participation in the seminars
- Be prepared to present your workflow diary and the current status of your project in each seminar session
- It is ok to not be able to present max. 2 times from Session 2 to Session 6

Assessment: Final Notebook

- Summarise the results of your analysis and workflow steps in the form of a case study
- Identify the results most relevant to your research question and present them in a clear and engaging way
- Emphasis on engagement with workflow tasks and transparency in the choices you made
- You will also peer review two notebooks from fellow students in the final week

Assessment: Requirements to pass the Course

Active participation in seminars

- Be present and ready to present your work in the seminars
- It is ok to not be able to present max. 2 times from Session 2 to Session 6
- Present results of your data analysis workflow in a final notebook

Continuous engagement with the topics covered each week

- Engage with workflow tasks for each week and apply them to your project
- Document your thoughts, observations and questions in your workflow diary
- Workflow diary has to have the content of each week and cover all the discussed workflow steps
- Submit workflow diary and notebook
- Complete peer review of other notebooks

Tools/Software

We will primarily provide resources for applying workflow steps in R & using Stan in the background

- Key Packages
 - Estimating models: `brms`, `cmdstanr`
 - Post-processing results: `posterior`, `tidybayes`
 - Diagnostic checks: `priorsense`, `loo`
 - Visualisations: `bayesplot`, `ggdist`

i Note

You are free to use any language, as long as you maintain a workflow diary and can interactively apply suggestions during presentations.

Plagiarism and AI tools

1.0.1 I have used AI for correcting my sentences

- Don't copy reports from others or from internet
- Mention the source for all materials and resources that you use
- It's OK to use AI, but you need to mention when and how it was used
 - Warning: AI tools can provide very vague or completely wrong results for the tasks in this course
 - Might be most useful for getting ideas for code and markdown syntax

Discussion of modelling problems & datasets

2 Minute Madness

Team up with the two persons next to you and **pitch your modelling ideas** to each other!

- Everyone gets 2min to prepare and 2min for their pitch
- Talk about a topic that you would like to investigate throughout the course
- What is the domain of this research?
- What kind of data would you use?
- If you already have a modelling problem or dataset, explain what you want to analyse and why

Open Discussion (Optional)

Let's take a moment and discuss some of the **modelling problems** and **datasets** that you are planning on working through!

- Opportunity to share your ideas
- Get initial feedback on what you want to analyse

Let's take a break! (10 min)

Some suggestions for recharging during breaks

- Move your body
- Open a window or go outside
- Drink some water
- Try to avoid checking e-mails, messengers, or social media

Primer: Choosing an Initial Model

For the next session, your tasks are

Starting the Project

- Formulating a research question & finding a dataset
- Visualising and getting familiar with characteristics of your data

Initial Model

- Picking an initial model & documenting your reasoning and the strategies you used to choose it
- Obtaining posterior samples using your initial model with default priors

Setup and Documentation

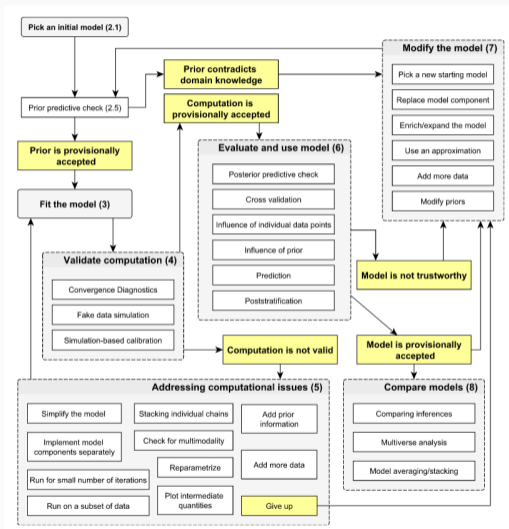
- Setting up your project, for example, using the provided templates
- Adding your notes and visualisations to the workflow diary
- Documenting observations and any issues you encounter in the workflow diary

Disclaimer

Note

Start simple! Your goal is not to identify all possible issues and propose the perfect model but to try your best to come up with a reasonable starting point. You will continue to refine your model and explore your data further throughout the course.

No Workflow without a Model



This week we are looking at how to pick an initial model. To support your work up until our next session, we will now talk about

- Modular model building
- Data model choice guided by research questions and characteristics of the data
- Strategies for picking an initial model

Building Models from Modules

We are interested in $p(\theta | y)$, for example, to obtain predictions

$p(\tilde{y} | y) = \int p(\tilde{y} | \theta) p(\theta | y) d\theta$ for new or future data \tilde{y}

$$p(\theta | y) \propto \underbrace{p(y | \theta)}_{\text{Data model}} \underbrace{p(\theta)}_{\text{Prior}}$$

Building Models from Modules

We are interested in $p(\theta | y)$, for example, to obtain predictions $p(\tilde{y} | y) = \int p(\tilde{y} | \theta) p(\theta | y) d\theta$ for new or future data \tilde{y}

$$p(\theta | y) \propto \underbrace{p(y | \theta)}_{\text{Data model}} \underbrace{p(\theta)}_{\text{Prior}}$$

We focus on $p(y | \theta)$ as a function of y (“data model”) for now:

- Identify outcome of interest
- Choose a **distributional family** for the observations y
- Decide which **predictors** to include and how to include them

Considering Research Questions

1. Decide on a specific research question and a dataset that contains relevant data for this question
2. Given the dataset, identify a research question that you can investigate with it
3. identify the outcome and other variables of interest from your research question

Research Questions - Example

“Does this drug treat depression?”

- Unclear what results would be needed to answer this
- Unclear how to construct a model to produce these

Compare this to ...

Research Questions - Example

“Do individuals taking the drug show greater reductions in depression over 4 weeks, compared to individuals that do not?”

- Clear outcome: Change in depression
- Clear structure: Comparing magnitude between groups
- Implies time-dependence of outcome (repeated measurements of depression)

Considering Characteristics of the Data

- How was the data collected?
- What scale is the outcome measured on? Discrete/Continuous/Skewed?
- What variables are included in your dataset?
- What are the measurement units?
- What ranges do your observations lie in?
- Are there interesting time-series properties (trends, seasonality, time-varying relationships)?

Using Templates, Literature and Recommendations

After having identified a clear and specific research question, picking an initial model can also be informed by

- Previous analyses with similar data in articles, tutorials or case studies
- Recommended models for this type of data or research question, for example, in textbooks
- Models that you are familiar with, for example, from previous courses
- A model that you already have and plan to extend considerably throughout the course

Resources for the Week

- Aki's talk "On Bayesian Workflow" at Bayes@Lund 2021 (<https://www.youtube.com/watch?v=lKRRyrPxxeU>)
- Draft of Bayesian Workflow book (not published yet, only made available for students in this course - please don't share!)
 - Chapter 1: Bayesian theory and Bayesian practice
 - Chapter 2: Computational tools
 - Chapter 3: Building a model
 - Chapter 15, Section 15.1: An example of a simple initial model for analysing sleep deprivation

Resources for the Week (ctd.)

- Introduction, Section 2.1 and 2.2 of “Bayesian Workflow” by Gelman et al. (2020) (<https://arxiv.org/abs/2011.01808>)
- Section 1-3 and 4.1 in Aki’s case study “Birthdays workflow example” (<https://users.aalto.fi/~ave/casestudies/Birthdays/birthdays.html>)

For the next session

- Check out the resources for the week
- Document your work in your workflow diary
- Prepare questions that you want to discuss

i Note

The clearer the questions and the more you experimented, the more you will get out of the discussion.

Getting started

Set up version control within RStudio

If using R and RStudio, R projects can be convenient for project setup and version control via GitHub.

Step 1: get the most recent versions of

- R (<https://www.r-project.org/>)
- RStudio (<https://posit.co/download/rstudio-desktop/>)
- Quarto (<https://quarto.org/docs/get-started/>)

Step 2: create a GitHub repository for your project

Step 3: create a new R project in RStudio and connect it to the remote repository




Step 4: add your data and Quarto template ([link](#)) to your R project

Step 5: install required packages

Set up version control within RStudio

New Project Wizard



Create Project

-  **New Directory**
Start a project in a brand new working directory >
-  **Existing Directory**
Associate a project with an existing working directory >
-  **Version Control**
Checkout a project from a version control repository >

Cancel

New Project Wizard

[Back](#) **Create Project from Version Control**

-  **Git**
Clone a project from a Git repository >
-  **Subversion**
Checkout a project from a Subversion repository >

Cancel

References

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. “Bayesian Workflow.” <https://arxiv.org/abs/2011.01808>.