Homework 4: Probabilistic Calibration and Model Selection

BEE 4850/5850, Fall 2024

Due Date

Friday, 5/03/24, 9:00pm

Tip

To do this assignment in Julia, you can find a Jupyter notebook with an appropriate environment in the homework's Github repository. Otherwise, you will be responsible for setting up an appropriate package environment in the language of your choosing. Make sure to include your name and NetID on your solution.

Overview

Instructions

The goal of this homework assignment is to practice probabilistic model calibration and using the resulting distributions for model evaluation and selection.

Learning Outcomes

After completing this assignments, students will be able to:

- quantify uncertainty in model parameters with probabilistic programming.
- use information criteria to assess the relative evidence for models.

Load Environment

The following code loads the environment and makes sure all needed packages are installed. This should be at the start of most Julia scripts.

```
import Pkg
Pkg.activate(@__DIR__)
Pkg.instantiate()
```

The following packages are included in the environment (to help you find other similar packages in other languages). The code below loads these packages for use in the subsequent notebook (the desired functionality for each package is commented next to the package).

```
using Random # random number generation and seed-setting
using DataFrames # tabular data structure
using DataFramesMeta # API which can simplify chains of DataFrames
 ↪  transformations
using CSVFiles # reads/writes .csv files
using Distributions # interface to work with probability distributions
using Plots # plotting library
using StatsBase # statistical quantities like mean, median, etc
using StatsPlots # some additional statistical plotting tools
```

Problems (Total: 30 Points)

Problem 1

This problem is the same as the MCMC Lab from earlier in this semester.

You will use a probabilistic programming language to fit a linear model for monthly mean tide gauge data from the Sewell's Point, VA tide gauge from 1927 through 2022, obtained from the Permanent Service for Mean Sea Level. The data (in `data/norfolk-monthly-tide-data.txt`) has been slightly cleaned by setting dates to the `yyyy-mm` format. We've left missing values as `-99999`; make sure to fix those as appropriate for your programming language.

```
tide_dat = CSV.read("data/norfolk-monthly-tide-data.txt", DataFrame)
# replace -99999 with missing
tide_dat.gauge = ifelse.(tide_dat.gauge .== -99999, missing, tide_dat.gauge)
```

```
LoadError: UndefVarError: `CSV` not defined
```

Now let's plot the data.

```
p = scatter(tide_dat.datetime, tide_dat.gauge, xlabel="Month",
 ↪  ylabel="Monthly Mean Sea Level (mm)", legend=false)
display(p)
```

```
LoadError: UndefVarError: `tide_dat` not defined
```

We would like to quantify the uncertainty in the time-trend of this local sea level increase (which includes global mean sea level rise but also more local effects, such as subsidence). The plot in looks roughly linear, so let's use the following model (assuming the errors are independent and identically-distributed for simplicity):

$$y(t) = \alpha + \beta t + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma).$$

In this problem:

- Write a model for the linear regression above in the probabilistic programming language of your choice. You'll need to pick some priors for α, β, and σ.
- Sample from the posterior with four chains (for convergence diagnostics).
- Evaluate convergence. How many iterations did you use? What is the effective sample size?
- Plot the posterior distributions. In particular, we are interested in uncertainty in the β coefficient, which reflects the mean increase in sea-level rise over time in mm/months.
- Generate hindcasts by sampling from the posterior distribution and simulating data. If you plot the 95% posterior predictive distribution and the data, how does it look?

Problem 2

Building on Problem 1, suppose we wanted to examine if the increase in sea level at Norfolk was quadratic instead of linear.

In this problem:

- Fit a quadratic model (with the same error structure) to the same data. What are the estimates?
- Plot the linear and quadratic fits along with the data. Do you visually see any differences?
- Compute the log-posterior, AIC, and DIC for each model. Based on these metrics, what would you conclude about the relative evidence for each model? What are your conclusions about whether the sea level trend is linear or quadratic?