1 What problem do your contributions aim to solve?

The ongoing pandemic COVID 19 has been impacting the entire world dramatically. This project aims to estimate and predict the mortality rate of 58 counties in California, using Bayesian approaches. With a very limited and noisy data, Bayesian approaches can not only provide point estimation, but also full probability models to quantify the uncertainty of the estimation.

1.1 What approach did you choose to solve this problem?

- Data Exploration
- Model Specification
- Sampling Importance Resampling (SIR) for Posterior Inference
- Predictive Posterior Inference
- Model Validation and Comparison

1.2 How does it work?

1.2.1 Dataset

The dataset I used contains the county population(c_i for i-th county), number of infected people(n_i), and number of death(y_i) due to COVID-19 in 58 counties of California, up to 04/13/2020,

1.2.2 Model Specification

I build two models to obtain inference on the number of deaths per 1000 residents (mortality rate) at each county of California. The two models differ by whether allowing infection rate depends on the mortality rate

- Assume the number of infection follows a Poisson distribution with certain rate and independent from mortality rate. Assume the mortality rate as a series of Bernoulli trials

$$\begin{aligned} &n_i \sim Pois(\lambda_i c_i/10^3), \lambda_i|\alpha,\beta \sim Gamma(\alpha,\beta), \alpha,\beta \sim Gamma \\ &y_i|n_i,\theta_i \sim Bin(n_i,\theta_i), \theta_i|\mu,\tau \sim Beta(\mu\tau,\tau(1-\mu)) \end{aligned} \quad [1.1]$$

- Assume the the infection rate depends on mortality rate:

$$\begin{aligned} &y_i|n_i,\theta_i \sim Bin(n_i,\theta_i), \\ &n_i|\theta_i,\lambda \sim Pois(\lambda_i\theta_i c_i/10^3) \end{aligned} \quad [1.2]$$

1.2.3 SIR Sampling for Posterior Inference

- For each model, I obtain 5000-10,000 posterior samples of the key parameters using "Sampling Importance Resampling" algorithm.
- The reason for obtaining posterior samples is to make sure that our models capture the real pattern of the data, so that we can get a better prediction result both in terms of accuracy and generalization capability.

1.2.4 Predictive Posterior Inference

- Obtain 90% percent interval estimates for the number of deaths per 1000 residents for each county.
- The estimated mean state-wise mortality rate up to 04/13/2020 is around 0.02 per 1000 people in California

1.2.5 Model Validation and Comparison

- Assess the validation of models by computing the distribution of Bayesian residual. If most counties have a Bayesian residual distributed around 0, then we can conclude that the model is valid as it captures the pattern well
- Use posterior predictive loss loss criteria to compute the goodness of fit of two models

2 In what context did you write this code?

It was a class project from the Advanced Bayesian Inference class. This is one part of the project

3 If this project was a collaboration

This is an individual project. No other people contributed to this project

4 What do you find interesting, surprising, or special about your contributions to this project?

It was at early stage of pandemic when I started this project. Though later on 0.02 mortality rate per 1000 residents was proved to be inaccurate, doing an analysis on covid-19 cases provided some objective insights into the possible direction California might head toward.

5 What did you learn from this project?

- I learned about the overall structure and workflow of a complete Bayesian analysis from EDA, to specifying priors to sampling from posterior distribution.
- To test whether our algorithm is reliable, one strategy is to fit the model to the simulated data to check if we capture the correct posterior distribution.

6 If you were to rewrite your code, what would you do differently?

I spent a lot of time building Bayesian models and sampling methods from scratch for learning purpose. Next time I would use built-in Bayesian packages such as Stan to save time and spend more time on analysis and organization of my codes