Bayesian Inference Final Project

Introduction

The reproduction number R_t is a common measure of transmissibility of an epidemic disease. By definition, R_t is the average number of secondary cases caused by an infected individual. R_t falling below 1 indicates the disease is unlikely to sustain. The change of R_t across time can be used as a proxy for epidemic trajectories. Following the method proposed by Cori et al., R_t can be calculated by using Bayesian parametric estimation. In this project, I use the same method to compute the R_t of COVID-19 in New York before the presence of effective vaccination (Mar 2020 to Dec 2020).

```
library(dplyr)
library(lubridate)
library(rstan)
```

```
df <-
  read.csv("NY_cases.csv")
df<-df%>%select(state,submission_date,new_case, tot_cases)
```

Method

Assuming the number of reported incident cases follows a poisson process, we have:

$P(I_t|I0, I1, .., I_{t-1}, w_s, R_t) = \frac{(R_t\Lambda_t(w_s))_t^I exp(-R_t\Lambda_t(w_s)}{I_t!}$,

where I_t is the number of incident cases arising at time t and $\Lambda_t(w_s)$ is the overall infectivity and can be computed as:

$\Lambda_t(w_s) = \sum_{s=1}^{t} I_{t-s}w_s$.

w_s is the serial interval distribution, which is the time between the onset of symptoms in a primary case and he onset of symptoms in secondary cases. It serves as the weight for reported new cases at each time step before time t. I choose the distribution of the serial period of COVID-19 follows a gamma distribution with mean of 5.9 and standard deviation of 3.9, according to the paper of Liu et al.. I also choose the prior of R_t to follow a gamma distribution with mean = 1.5 and standard deviation of 2.

With the above assumptions and procedure, I use Stan to yield the posterior distribution of R_t on the last day of 2020 (Dec 31st, 2020) with all previous reported cases.

```
#get the weight of previous days using a gamma distribution
lastday <- df %>%
  mutate(days_away=rev(row_number())-1,
         weight=dgamma(days_away,shape=(5.9/3.9)^2,rate=5.9/3.9^2))
```

```
data { /* these are known and passed as a named list from R */
  int<lower = 0> I;             // number of cases in day t
  real<lower = 0> infect;  // weighted sum of previous infectivity
  real<lower = 0> alpha; //shape parameter of gamma prior
```

```
  real<lower =  0> beta;// rate parameter of gamma prior
  int<lower = 0, upper = 1> prior_only;
}

parameters {
  real<lower=0> Rt; // Reproduction number
}

model {
if (!prior_only) {
  target +=poisson_lpmf(I | Rt*infect); // log-likelihood
  }
  target += gamma_lpdf(Rt | alpha, beta); //prior of Rt
}
```

```
#calculating overall infectivity
overallinfectivity=sum(lastday$new_case*lastday$weight)
#indicate new case on Dec 31,2020
It=lastday$new_case[nrow(lastday)]
#choose prior parameters (shape and rate)
a=9/16
b=3/8
```

```
post <- stan("reproduction_num.stan",
             data = list(infect=overallinfectivity,I=It, prior_only = 0, alpha = a, beta=b))
```

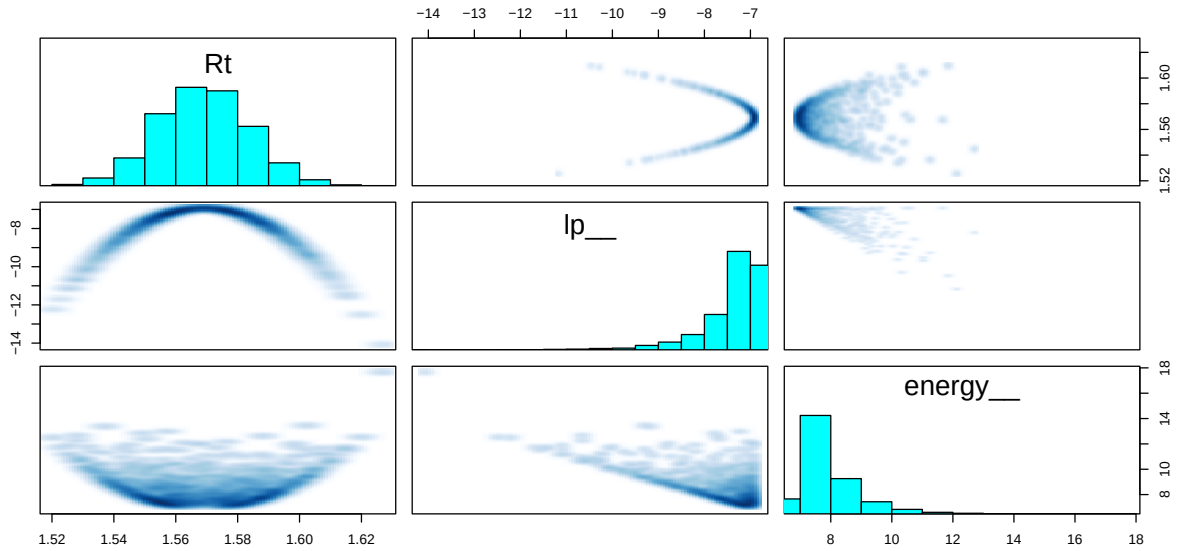
```
post
```

```
## Inference for Stan model: reproduction_num.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##       mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## Rt    1.57    0.00 0.01  1.54  1.56  1.57  1.58  1.60  1473    1
## lp__ -7.38    0.01 0.66 -9.29 -7.55 -7.13 -6.96 -6.91  1923    1
##
## Samples were drawn using NUTS(diag_e) at Sun May 15 21:53:15 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

The mean of the posterior distribution of R_t on Dec.31st, 2020 in New York is 1.57 with a 0.95 credible interval of (1.54,1.60).

Posterior Planes

```
pairs(post, pars = "p", include = FALSE)
```




```
draws <- as.data.frame(post) %>% select(-starts_with("p"))
mean(draws$Rt)
```

```
## [1] 1.568948
```

Different Assumptions

I then modify the approach by using a negative binomial likelihood instead of a poisson likelihood since negative binomial distribution allows for overdispersion. The negative binomial distribution has an extra dispersion parameter ϕ and it is assumed to follow a half cauchy distribution with location parameter =5 and scale = 5.

```
data { /* these are known and passed as a named list from R */
  int<lower = 0> I;              // number of cases in day t
  real<lower = 0> infect;  // weighted sum of previous infectivity
  real<lower = 0> alpha; //shape parameter of gamma prior
  real<lower =  0> beta;// rate parameter of gamma prior
  real<lower = 0> mu; //location parameter of cauchy prior
  real<lower = 0> sigma;//scale parameter of cauchy prior
  int<lower = 0, upper = 1> prior_only;
}

parameters {
  real<lower=0> Rt; // Reproduction number
  real<lower=0> phi; // dispersion parameter
}



model {
if (!prior_only) {
  target +=neg_binomial_2_lpmf(I | Rt*infect, phi); // log-likelihood
```

```
  }
  target += gamma_lpdf(Rt | alpha, beta); //prior of Rt
  target += cauchy_lpdf(phi | mu, sigma); //prior of phi
}
```

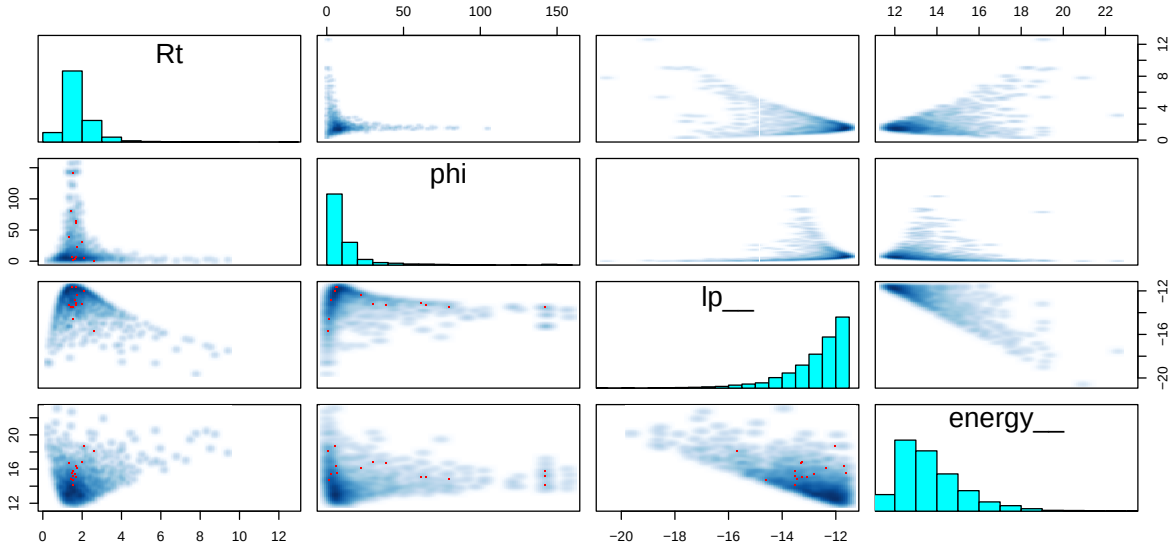
```
post_neg <- stan("reproduction_num2.stan",
             data = list(infect=overallinfectivity,I=It, prior_only = 0,
                         alpha = a, beta=b, mu=5,sigma=5))
```

```
post_neg
```

```
## Inference for Stan model: reproduction_num2.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##         mean se_mean    sd    2.5%    25%    50%    75%  97.5% n_eff Rhat
## Rt      1.79    0.03  0.88    0.77   1.29   1.60   2.04   3.96  1083 1.00
## phi    12.73    1.05 18.85    1.00   4.41   7.37  12.46  66.75   324 1.01
## lp__ -12.73    0.04  1.17 -15.92 -13.20 -12.37 -11.88 -11.54   830 1.01
##
## Samples were drawn using NUTS(diag_e) at Sun May 15 21:53:38 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

After modification, the negative binomial likelihood gives a wider credible interval which implies higher uncertainty in estimation.

```
pairs(post_neg, pars = "p", include = FALSE)
```



Essentially, we can compute the posterior distribution of R_t for all time points in the data set,assuming the same serial interval distribution. Using the mean of the posterior distributions, a graph of R_t can be plotted.

```
for (i in 2:nrow(df)){
  dt<-df[1:i,] %>%
  mutate(days_away=rev(row_number())-1,
         weight=dgamma(days_away,shape=5.9^2/3.9,rate=5.9/3.9))
  overall=sum(dt$new_case*dt$weight)
  I_t=dt$new_case[nrow(dt)]
  a=9/16
  b=3/8
  posterior <- stan("reproduction_num.stan",
              data = list(infect=overall,I=I_t, prior_only = 0, alpha = a, beta=b))
post.df <- as.data.frame(posterior) %>% select(-starts_with("p"))
df$Rt[i]<-mean(post.df$Rt)
}
```

Reference

Anne Cori, Neil M. Ferguson, Christophe Fraser, Simon Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, American Journal of Epidemiology, Volume 178, Issue 9, 1 November 2013, Pages 1505–1512, https://doi.org/10.1093/aje/kwt133

Liu, X., Xu, X., Li, G., Xu, X., Sun, Y., Wang, F., Shi, X., Li, X., Xie, G., & Zhang, L. (2021). Differential impact of non-pharmaceutical public health interventions on COVID-19 epidemics in the United States. BMC Public Health, 21(1), 965. https://doi.org/10.1186/s12889-021-10950-2