Parameter Inference on a Modified SIR Model for Disease Spread

Gregory Orme

1 Introduction

I originally started on this project during summer 2021 as a student researcher at Duke University. Dr. David Dudson, along with doctoral student Omar Melikechi and postdoc Tao Tang, were drafting a paper, "Limits of epidemic prediction using SIR models" (2022), investigating difficulties with parameter inference in vanilla SIR (susceptible, infected, recovered) models of disease spread. As a student researcher that summer, I explored an alternative SIR framework for epidemic prediction that allows for robust parameter inference using live data. **My research goals in this project were twofold: first, to apply Bayesian methods to overcome structural barriers to parameter inference, and second, to restate the vanilla SIR model in a format that does not rely on new case counts to make inferences**. Early in an epidemic, new case counts are likely to understate the extent of disease spread due to lack of testing resources, potentially introducing model bias toward lower projections of future case counts, hospitalizations, and deaths.

I recently revisited my original work on the problem, producing the following model. This paper will cover the conceptual background, model dynamics, parameter inference scheme, implementation in Python, results from simulations on a real world dataset, and finally some areas for improvement.

2 SIR Models and Problem Statement

The original SIR model is a simple and widely known compartmental model of disease spread. Compartmental models divide a vulnerable population into subsets whose relative sizes evolve according to differential equations prescribed by the model. For example, the original SIR model assumes all members of the population are either susceptible to the disease, currently infected, or recovered, hence the SIR acronym. After fixing an initial allocation of the three compartments (S_0, I_0, R_0), the model uses three differential equations with two additional parameters to describe the future states of the disease. The equations are as follows:

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

Here, S is the susceptible fraction of the population, I is the infected fraction, R is the recovered fraction, t is time, β is the expected number of new infections caused by each infected person per unit time, γ is the percent of infected people who should recover each day, and $S + I + R = 1$ (conservation

of population). In simpler terms, $\frac{1}{\gamma}$ is the expected time from becoming contagious to becoming noncontagious, β is the rate of spread, and the ratio $\frac{\beta}{\gamma}$ is the reproduction number (total expected new infections caused by each infected person).

In "Limits of epidemic prediction using SIR models" (2022), Dunson et al. showed that the vanilla SIR model above produces segments of points in (β, γ) space with nearly indistinguishable paths for the infected population in the early period of the disease. That is, if an epidemiologist knew the recent path of COVID infections perfectly and wanted to recover the parameters (β, γ) to predict the forward evolution of the disease, there would be infinitely many (β, γ) pairs matching the observed data. Without additional features, the SIR model is opaque to parameter inference, making it a blunt tool for disease prediction. To address this problem, I will develop a Bayesian approach to leverage our prior knowledge of disease spread parameters. **This is my first research goal stated in the introduction.**

To update those prior beliefs, the model must integrate outside information about the spread of disease in real time. My naive first approach was to identify official new case counts with the βSI term in (1) and (2). Then, assuming some amount of noise in case reporting, I implemented a simple likelihood function to model the probability of an observed sequence of new cases given any pair of parameters (β, γ).

This approach works only if we believe that new case counts reflect the true extent of new infections. However, testing availability may significantly lag the spread of disease early in a breakout. Even if tests are available, large numbers of asymptomatic or moderately symptomatic carriers may opt not to test. Both of these factors were present early in the COVID-19 pandemic in the United States, where retrospective estimates of true case counts have been much higher than the confirmed numbers (for an example estimate, see Dr. Jungsik Noh's work on the subject at https://github.com/JungsikNoh/COVID19_Estimated-Size-of-Infectious-Population).

Backing out the ratio of true cases to reported cases is not a straightforward problem. Even if I could model that relationship confidently, the SIR dynamics still require an initial condition (S_0, I_0) which is similarly challenging to estimate. Instead of using new cases, I want to use an alternative indicator of disease spread whose relationship with the unknown infected population is easier to model. Then, after revising the SIR model to reflect that relationship, I will derive a new likelihood function to mediate parameter inference. **Reworking SIR to accept this alternative data approach is my second research goal stated in the introduction.**

3 Model

Instead of using (1)-(3), I will use a slightly modified SIR model, SIRH: susceptible, infected, recovered, and hospitalized. Dynamics are as follows:

$$\frac{dS}{dt} = -\rho\gamma SI \tag{4}$$

$$\frac{dI}{dt} = -\rho\gamma SI - (1+\phi)\gamma I \tag{5}$$

$$\frac{dR}{dt} = \gamma I \tag{6}$$

$$\frac{dH}{dt} = \phi\gamma I - \gamma H \tag{7}$$

For some initial condition (S_0, I_0, R_0, H_0),

where S, I, R, γ are the same as in vanilla SIR, H denotes the hospitalized fraction of the population, ρ is the reproduction rate of the disease ($\frac{\beta}{\gamma}$), and ϕ is the probability that an infected person will become hospitalized during the entirety of their illness. If ϕ is the hospitalization probability per case, then $\phi\gamma$ signifies the hospitalization probability per case per day (exercise for the reader if curious), giving rise to the $\phi\gamma I$ terms in (5) and (7).

I replace β with a reproduction number term ρ because it is easier to conceive prior beliefs about the reproduction number than about the expected number of new infections per day caused by each infected person. Estimating new infections per day per infected person requires modelling the average number of daily social contacts per person as well as the probability of infection per social contact. On the other hand, estimating the reproduction number entails calculating simple averages from contact tracing networks. At the very least, contact tracing data can put effective lower bounds on the reproduction number.

The motivation for adding a hospitalized compartment is the fact that hospitalization data should reflect the size of the infected population more accurately than new cases or death counts. Especially in the early phase of an epidemic, case counts are likely to underestimate the true extent of new cases, while death counts tend to exclude victims who succumb outside of the hospital system (such as in nursing homes). Hospitalization counts, at least in major cities with testing resources, should miss very few people under hospital care with disease symptoms. Therefore, if we have prior beliefs about rates of severe illness resulting in hospitalization, we can use new hospitalization data to back out the infected population over some trailing period.

Even once we glean the recent trajectory of the infected population with some confidence, the parameter inference problem remains. My SIRH model is a slightly embroidered SIR model; the point of adding the H component is to improve inference on the infected population, not the parameters. Therefore, we need an inference mechanism to transform our prior beliefs about the parameters into estimates of their true values, using hospitalization data as an avenue to model the trend in the susceptible and infected populations.

4 Likelihood Function

I will split my sequence of new hospitalizations by day into blocks of predetermined size. In the example later on, I use 15 days. Given a sequence of observed new hospitalizations H_{new} and any five-tuple $(\rho, \gamma, \phi, S_0, I_0)$, where S_0 and I_0 are the susceptible and infected fractions at the start of the 15 day block, I aim to derive a probability of observing this sequence of hospitalizations assuming that the compartments $(S(t), I(t), R(t), H(t)$ evolve according to the dynamics in equations 4-7. That is, I want to find a likelihood function $L(H_{new}|(\rho, \gamma, \phi, S_0, I_0))$.

First, a quick note on why I exclude initial conditions R_0 and H_0 from the likelihood function. I am treating $H(t)$ and $R(t)$ as a block removed population, which makes sense because under most circumstances $R(t) >> H(t)$. So I let S_0 and I_0 implicitly specify $H_0 + R_0$ through the identity $S + I + R + H = 1$. Moreover, the count of new hospitalizations on time t has no dependence on the existing hospitalizations $H(t-1)$, rendering H_0 an unnecessary detail to account for separately.

The inputs (ρ, γ, S_0, I_0) imply deterministic

paths for $S(t)$ and $I(t)$, from which ϕ and γ specify a path for new hospitalizations (and hence our observed data H_{new}). If an infected person has a $\gamma\phi$ probability of becoming hospitalized each day, then at time t we can model the total new hospitalizations as a binomial distribution with mean $I(t)\gamma\phi$ and variance $I(t)\gamma\phi(1-\gamma\phi)$. Let $I(t)$ be the deterministic path for the infected population specified by some five tuple $(\rho, \gamma, \phi, S_0, I_0)$. Then our likelihood function is:

$$\mathrm{L}(\mathrm{H}_{new}|(\rho,\gamma,\phi,S_0,I_0)) = \prod_{k=1}^{len(H_{new})} \frac{1}{\sqrt{2\pi I(k)\gamma\phi(1-\gamma\phi)}} e^{\frac{-(H_{new}(k)-I(k)\gamma\phi)^2}{2I(k)\gamma\phi(1-\gamma\phi)}} \quad (8)$$

Each day k, I am calculating the probability of observing $H_{new}(k)$ assuming new hospitalizations are binomially (normally) distributed about $I(k)\phi\gamma$, where $I(k)$ is our deterministic total infected population for day K derived from SIRH dynamics. Noting that the observations are independent, I can multiply these probabilities consecutively to derive the total probability of observing this exact sequence H_{new} assuming parameters $(\rho, \gamma, \phi, S_0, I_0)$, justifying our likelihood function in (8). In practice $L(H_{new}|(\rho, \gamma, \phi, S_0, I_0))$ will be very close to 0 due to repeated multiplications of probabilities < 1, so I take logs instead:

$$\log L(H_{new}|(\rho,\gamma,\phi,S_0,I_0)) = \sum_{k=1}^{len(H_{new})} (\log \frac{1}{\sqrt{2\pi I(k)\gamma\phi(1-\gamma\phi)}} + \log \frac{-(H_{new}(k)-I(k)\gamma\phi)^2}{2I(k)\gamma\phi(1-\gamma\phi)}) \quad (9)$$

(9) leverages the fact that logs turn products into sums. Lastly, we need to attach priors to some subset of parameters in $(\rho, \gamma, \phi, S_0, I_0)$ to disambiguate (ρ, γ) pairs that specify similar infection paths $I(t)$. Let π_λ denote a prior function for an arbitrary parameter λ, and let $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ be a five tuple of parameters. Let $\xi(H_{new}, \rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ be the posterior distribution of the parameters $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ given observation H_{new}. Using Bayes' rule, $\log \xi$ becomes (removing integration constants and other terms constant across all arguments):

$$\log \xi(H_{new}, \rho^*, \gamma^*, \phi^*, S_0^*, I_0^*) = \sum_{k=1}^{len(H_{new})} (\log \frac{1}{\sqrt{2\pi I(k)\gamma^*\phi^*(1-\gamma^*\phi^*)}} + \log \frac{-(H_{new}(k)-I(k)\gamma^*\phi^*)^2}{2I(k)\gamma^*\phi^*(1-\gamma^*\phi^*)}) + \log \pi_{S_0}(S_0^*) + \log \pi_{I_0}(I_0^*) + \log \pi_\gamma(\gamma^*) + \log \pi_\phi(\phi^*) + \log \pi_\rho(\rho^*) \quad (10)$$

5 Parameter Inference

Equipped with the log posterior function $\log \xi$ in (10), we can apply an iterative Markov chain algorithm to explore and sample parameters. I implement a simple Metropolis-Hastings algorithm with normal jump distributions for (γ, ϕ, ρ) with fixed standard deviations for the proposals. More concretely, for (γ, ϕ, ρ), I draw new proposals from normals centered at 0 with standard deviation proportional to the standard deviation of our prior for those parameters (ex. one eighth the standard deviation of the prior). For S_0 and I_0, the jump distribution is normal with standard deviation proportional to the distance of the current S_0 or I_0 from boundaries 0 and 1. For example, if current S_0 is 0.1 and we are using a 2% standard deviation to propose the next S_0, we sample from a normal with mean zero and standard deviation 0.002 $(0.1 * 0.02)$ and add that number to our current $S_0 = 0.1$. Conversely, if the current S_0 is 0.95, we sample from a normal with mean zero and standard deviation 0.001 $(0.05 * 0.02)$ and add to our original $S_0 = 0.95$.

I use fixed proposal standard deviations for (γ, ϕ, ρ) because our prior estimates for those parameters are unlikely to diverge from the true values by orders of magnitude. On the other hand, (S_0, I_0) can range from a few dozen people to millions. As a result, I vary the standard deviation of proposals for (S_0, I_0) so that proposals remain proportional to the current values. This kind of relative sampling is important at the beginning of an epidemic when $I \approx 0$ and $S \approx 1$, ensuring that proposals remain in context of the appropriate order of magnitude.

Exact details on the Metropolis- Hastings step are visible in the function *mcmc* in the commented code, attached at https://github.com/gregorme2001/SIR-Model/blob/main/SIR_revised.ipynb

When updating priors, I treat (γ, ϕ, ρ) differently from (S_0, I_0). After running *mcmc*, I reset prior means and standard deviations for (γ, ϕ, ρ) to the posterior means and standard deviations. For (S_0, I_0), I do not use posterior estimates to inform prior beliefs for the next run, though I do use the function *project_s_i* explained in the next section to calculate first estimates for (S_0, I_0) in the next run of parameter inference. I want to give the Metropolis-Hastings process as much freedom as possible to explore values for (S_0, I_0), because (S_0, I_0) are the parameters for which we have the least informative prior beliefs at any time. Therefore, aside from initial estimates informed from our last run of the simulation, I impose uniform priors on (S_0, I_0) at all times. In practice, our priors on ϕ and γ implicitly force prior beliefs on I_0, but I don't want to constrain I_0 even further with potentially ill-informed priors.

6 Code Implementation

For the code itself, please see the link in section "Code Appendix" at the end.

I define a class *epidemic* to store our hospitalization data, prior beliefs, numerical integrator, likelihood function, Metropolis-Hastings algorithm, prediction functions, and the iterative updating scheme. The class *epidemic* takes the following arguments:

- *guess_gamma*: our initial guess for the parameter γ
- *guess_repro_rate*: our initial guess for the parameter ρ
- *guess_phi*: our initial guess for the parameter ϕ
- *guess_susceptible*: our guess for the number of people susceptible on the day of the first reported hospitalization
- *guess_infected*: our guess for the number of people infected on the day of the first reported hospitalization
- *gamma_stdv*: standard dev of our first prior on γ
- *repro_rate_stdv*: standard dev of our first prior on ρ
- *phi_stdv*: standard dev of our first prior on ϕ
- *s_0_stdv*: governs standard dev of proposals for S_0 in the mcmc step. I set uniform priors on S_0 in my runs later on
- *i_0_stdv*: governs standard dev of proposals for I_0 in the mcmc step. I set uniform priors for I_0 in my runs later on
- *population*: total population of the region under study
- *hospitalizations*: list to store total new hospitalizations on each day. This data guides our parameter search
- *mcmc_lookback*: how many days in the past *mcmc* should use data for parameter search. So if $mcmc\_lookback = 15$, we use 15 days of trailing hospitalization data to infer parameters
- $freq$: how frequently we rerun the parameter inference process. If $freq = 5$, then every 5 days, we run parameter inference and update our priors
- *max_days*: how many days of data the algorithm should run on. If $max\_days = 100$, then once the update function sees the hundredth day of data, the update process terminates.

Next, a summary of relevant functions and their purpose:

- *gamma_prior*: prior for the parameter γ
- *repro_rate_prior*: prior for the parameter ρ
- *phi_prior*: prior for the parameter ϕ
- *integrate_from_initial_conditions*: given a five tuple $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ and a number of days *days*, calculate and return the deterministic paths for $S(t)$ and $I(t)$ based on these parameters out to time $t = days$
- *log_likelihood*: given a five tuple $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ and a list of *hospitalization_data*, calculate the log posterior function $\log \xi$ defined in (10)
- *mcmc*: given a list of *hospitalization_data*, and using our current priors, explore the posterior using Metropolis-Hastings and return accepted samples for $(\rho, \gamma, \phi, S_0, I_0)$. Perform 500k iterations (arbitrary) and shed first quarter of accepted samples as burn in
- *project_s_i*: given a list of accepted five tuples of the form $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$ and a target date *target_index*, sample from the list and project the tuple $(S_{target\_index}, I_{target\_index})$ for each sample using *integrate_from_initial_conditions*.

That is, sample from the distribution of the values of S and I at the future date $target\_index$

- $project\_path$: given a list of accepted five tuples of the form $(\rho^*, \gamma^*, \phi^*, S_0^*, I_0^*)$, use the numerical integrator $integrate\_from\_initial\_conditions$ to calculate predicted curves for the infected population out to 150 days. Return a 95% confidence interval on the date of the peak and plot 95% confidence intervals for the infected population over the next 150 days.
- $update$: move 1 day in the future. If the number of days since last update equals $freq$, rerun parameter inference using $mcmc\_lookback$ days of hospitalization data. Print 95% confidence intervals on the values of interest, update prior beliefs, and return.
- $run\_simulation$: Run the function $update$ every $freq$ days until $day = max\_days$.

7 Data

I use daily hospitalization counts reported by New York City in the csv "COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths_20240630.csv" available at https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3/about_data. The column "HOSPITALIZED_COUNT" reports the number of new hospitalizations per day across the entire city.

8 Results

Below, I have run the simulation process for forty days on hospitalization data for New York City at the beginning of the pandemic. Day 0 is the day on which NYC reported its first hospitalized patient. Parameters of the class epidemic are as follows:

- $guess\_gamma$: 0.1 (assuming 10 days to recovery on average). In a live setting, a public health expert would choose this parameter along with the other guess arguments that control our prior beliefs.
- $guess\_repro\_rate$: 4. Assuming that the average infected person infects four others
- $guess\_phi$: 0.03. Assuming 3% chance of hospitalization per case
- $guess\_susceptible$: Assuming 7.5 million have no immunity whatsoever. Pure shot in the dark
- $guess\_infected$: 10000. Guessing 10k active infections in the city at the time of first hospitalization reported. Also a shot in the dark.
- $gamma\_stdv$: 0.02. Implies we are 95% confident recovery takes between 7 and 17 days (noting that $\frac{1}{\gamma}$ gives the expected time to recovery, so our interval is $(\frac{1}{0.14}, \frac{1}{0.06}) \approx (7, 17)$ (this would be informed by public health expert at time)
- $repro\_rate\_stdv$: 1. 95% confident that the average person infects 2 to 6 other people (this would be informed by public health expert at time)
- phi_stdv: 0.01. 95% confident the hospitalization rate is between 1% and 5%
- s_0_stdv: 0.25%. Arbitrary but works in practice. Does not affect priors/posteriors since I use uniform prior beliefs on the number of susceptible/infected people to start. This parameter controls the standard deviation of the proposal distribution for S_0 during the $mcmc$ step
- i_0_stdv: 0.25%. Arbitrary but works in practice. Same as S_0 above
- $population$: 8.5 million. Population NYC

- *hospitalizations*: From https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3/about_data
- *mcmc_lookback*: 15 days. At the discretion of user but I think good window
- *freq*: 5 days. Rerun parameter inference every 5 days
- *max_days*: 40 days. New hospitalizations peak at day 30 so we run slightly through that date.

Now, screenshots of the console output:

```
Day: 0, not enough data to begin simulation
Day: 1, not enough data to begin simulation
Day: 2, not enough data to begin simulation
Day: 3, not enough data to begin simulation
Day: 4, not enough data to begin simulation
Day: 5, not enough data to begin simulation
Day: 6, not enough data to begin simulation
Day: 7, not enough data to begin simulation
Day: 8, not enough data to begin simulation
Day: 9, not enough data to begin simulation
Day: 10, not enough data to begin simulation
Day: 11, not enough data to begin simulation
Day: 12, not enough data to begin simulation
Day: 13, not enough data to begin simulation
Day: 14, not enough data to begin simulation
```

Figure 1: Output during first 15 days: not enough data in lookback window to run simulation.

After 15 days, the lookback window equals parameter *mcmc_lookback*, allowing parameter inference to run:

```
Day:  15
Rho 95% confidence interval:  (3.7798, 6.1531)
Gamma 95% confidence interval:  (0.0724, 0.1303)
Phi 95% confidence interval:  (0.0061, 0.0272)
Currently susceptible 95% confidence interval:  (0.7056, 0.9839)
Currently infected 95% confidence interval:  (0.0079, 0.0289)
Peak time 95% confidence interval:  (27, 34)
```

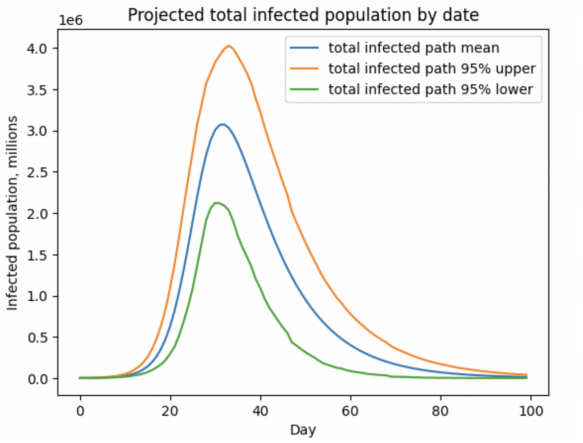

Figure 2: Output at day 15 after first hospitalization (using days $0-14$ of hospitalization data)

The inference tool prints the posterior 95% confidence intervals for ρ, γ, and ϕ during the first 15 days since the first reported hospitalization. The tool also uses the list of accepted five tuples $(\rho, \gamma, \phi, S_0, I_0)$ and the system of equations (4)-(7) to provide 95% confidence intervals on the fraction of the population that is currently susceptible/infected. Also using equations (4)-(7), the tool prints a 95% confidence interval on the day when the infected population will peak. Finally, the tool charts an expected path for the infected population alongside 95% upper and lower bounds for that path. (console output continues next page)

```
Day:  20
Rho 95% confidence interval:  (4.1096, 6.2721)
Gamma 95% confidence interval:  (0.0821, 0.1344)
Phi 95% confidence interval:  (0.0056, 0.0233)
Currently susceptible 95% confidence interval:  (0.4624, 0.8117)
Currently infected 95% confidence interval:  (0.0281, 0.1)
Peak time 95% confidence interval:  (28, 35)
```

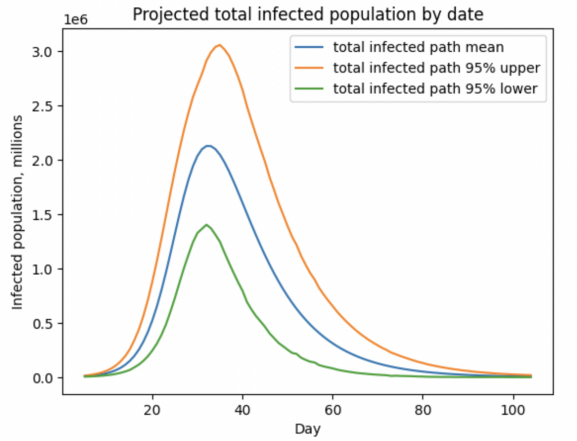

Figure 3: Output at day 20 after first hospitalization (using days $5 - 19$ of hospitalization data)

```
Day:  25
Rho 95% confidence interval:  (4.3895, 6.1405)
Gamma 95% confidence interval:  (0.0848, 0.1298)
Phi 95% confidence interval:  (0.0061, 0.012)
Currently susceptible 95% confidence interval:  (0.2931, 0.4491)
Currently infected 95% confidence interval:  (0.1058, 0.2182)
Peak time 95% confidence interval:  (29, 33)
```

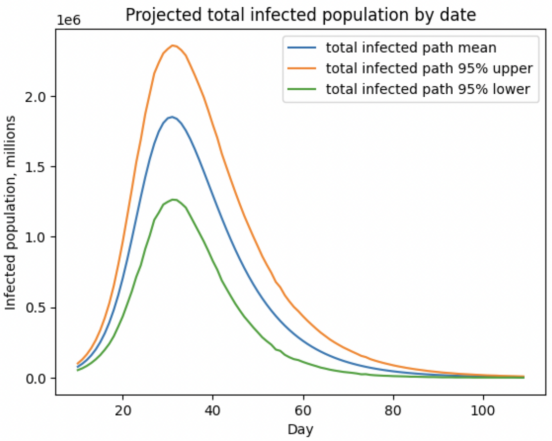

Figure 4: Output at day 25 after first hospitalization (using days $10 - 24$ of hospitalization data)

```
Day:  30
Rho 95% confidence interval:  (4.7769, 6.4076)
Gamma 95% confidence interval:  (0.1001, 0.1419)
Phi 95% confidence interval:  (0.0071, 0.0105)
Currently susceptible 95% confidence interval:  (0.1472, 0.2076)
Currently infected 95% confidence interval:  (0.1179, 0.2231)
Peak time 95% confidence interval:  (29, 30)
```

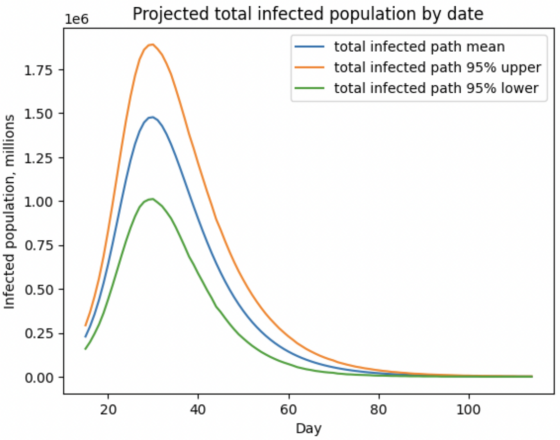

Figure 5: Output at day 30 after first hospitalization (using days $15 - 29$ of hospitalization data)

The algorithm provides sensible ranges for γ, ρ, ϕ, and peak time that are either stationary or continuously trending. The prior 95% confidence interval for ρ before any inference is

```
Day:  35
Rho 95% confidence interval:  (4.3846, 5.7876)
Gamma 95% confidence interval:  (0.0885, 0.1284)
Phi 95% confidence interval:  (0.0088, 0.0115)
Currently susceptible 95% confidence interval:  (0.1153, 0.171)
Currently infected 95% confidence interval:  (0.1369, 0.233)
Peak time 95% confidence interval:  (31, 32)
```

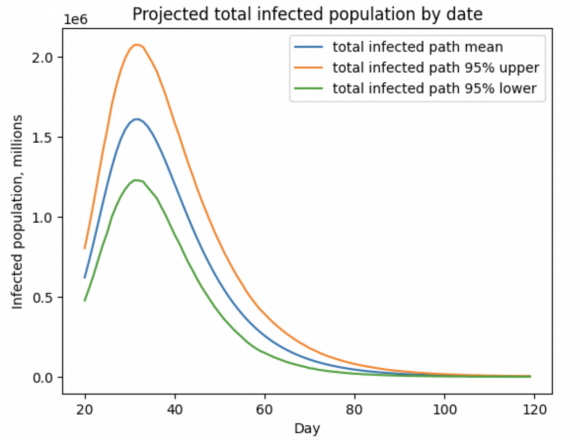

Figure 6: Output at day 35 after first hospitalization (using days $20-34$ of hospitalization data)

$(2, 6)$; at the first run on day 15, the algorithm revises upward to $(3.75, 6.15)$, and by the final run on day 35 has the interval at $(4.4, 5.8)$. γ remains stable throughout in the $(0.07, 0.13)$ range, while ϕ starts at $(0.01, 0.05)$ and gets revised down to $(0.009, 0.012)$. I deliberately set the prior mean on ϕ to be higher than later estimates, because reports from the COVID epidemic in Italy at the time were putting hospitalization rates as high as the mid single digits. Estimates for the peak date of the infected population also remain stable in the range of days $27-35$, while the observed peak date for hospitalizations in our dataset is 30 (the algorithm does not know day 30 is the empirical peak until day 35, when the trailing data window includes days $20-34$). On day 35, the algorithm knows the epidemic has peaked, reporting a peak time 95% interval of $(31, 32)$.

Projections of the peak infected population, on the other hand, start at $2-4$ million on day 15 and continuously revise downward to $1-2$ million on day 35. Even after the epidemic has peaked on day 30, the algorithm is not good at retrospectively pinpointing how many people were infected at the peak. This uncertainty justifies my apprehension about putting any priors on the initial conditions (S_0, I_0), because even posterior estimates are unacceptably wide.

9 Conclusion, Limitations, and Improvements

The SIRH model I outlined in equations $(4)-(7)$ does not solve the parameter inference problem on its own, instead creating a fourth compartment (hospitalized population) that is easier to observe. To deal with parameter non-identifiability, I set priors on the parameter values and iteratively update with a Metropolis-Hastings process. Using COVID hospitalization data from March-April 2020 in NYC, I run simulations every five days to infer my disease spread parameters γ, ϕ, ρ, my initial conditions (S_0, I_0), the future peak date of the infected population, and the number of people infected at the peak. The model provides accurate forecasts of the peak date but lacks precision on the number of people who will be infected at the peak.

A skeptical reader might also question how we can arrive at priors for γ, ϕ, ρ in the first place near the beginning of an epidemic, especially as these parameters might be population specific. My run above imposes wide priors on the spread parameters (95% range for the hospitalization rate ϕ is $(1\%, 5\%)$, 95% range for the number of days to recovery $\frac{1}{\gamma}$ is $(7, 17)$, 95% range for the reproduction number ρ is $(2, 6)$. The parameter inference process gradually refines my prior beliefs for ρ and ϕ, pinning ρ in the

$(4.4, 5.8)$ range and ϕ in the $(0.9\%, 1.3\%)$ range.

The most compelling challenge to this methodology is the size of the five-dimensional parameter space $(S_0, I_0, \gamma, \phi, \rho)$. In 500k iterations, is the algorithm really exploring the posterior properly? Are there disjoint regions of high posterior mass which are unexplorable without large jumps? The answer to this question requires more digging into the system $(4) - (7)$. In Dunson et al. (2022), the authors found early-epidemic nonidentifiability along lines of slope 1 in (β, γ) space. Setting priors on γ and $\rho = \frac{\beta}{\gamma}$ should eliminate those likelihood ridges locally but does not preclude the existence of disjoint likelihood peaks.

10 Code Appendix

https://github.com/gregorme2001/SIR-Model/blob/main/SIR_revised.ipynb

11 Citations

- Melikechi O, Young AL, Tang T, Bowman T, Dunson D, Johndrow J. Limits of epidemic prediction using SIR models. J Math Biol. 2022 Sep 20;85(4):36. doi: 10.1007/s00285-022-01804-5. PMID: 36125562; PMCID: PMC9487859.