Bayesian ODE/Convolution Models for Estimating Underlying Growth of COVID-19 and its Uncertainty

Douglas Mason[*], Robert Martinez[†]

May 8, 2020

We model universal curves of reported COVID-19 daily reported infections and related deaths using a modified epidemiological Susceptible-Exposed-Infectious-Recovered (SEIR) Model[6, 2, 7]. Using currently available data, we determine optimized constants and apply this framework to reproducing the infection and death curves for California (the state with the largest population), New York (the state with highest population density), and U.S. totals, and provide supplementary results for the remaining 50 states and Washington D.C. Source code used to produce these results can be found at the companion website[3]. Data is sourced from the New York Times[8].

1 Model Definition

It is helpful to define various sets that appear in the model as time-dependent variables. In the early stages of an epidemic or pandemic, the vast majority of individuals are susceptible to infection, while few individuals have recovered from an infection, and for this reason we can approximate the S and R values to 100% and 0% respectively. In addition, the SEIR model treats transitions from one group to the next as emissions with fixed (or possibly time-varying rates), which may be accurate over long time periods, but does not reflect the dynamics we attempt to model. Rather, while the growth of contagious individuals is well-described by time-varying growth rates (which can be solved by integrating a single ordinary differential equation), the transition to other measurable states (such as being tested and confirmed positive, or subsequently dying) likely exhibits an average time-delay with a measurable variance and an overall multiplier. For example, a person who has just become contagious will likely not show symptoms for up to a week, and then show symptoms, and then go to the hospital and be tested, and then have a probability of being confirmed positive.

[*] Koyote Science LLC and Nexus iR&D Laboratory, San Francisco, CA

[†] Harvard University, Cambridge College, and Nexus iR&D Laboratory, Boston, MA

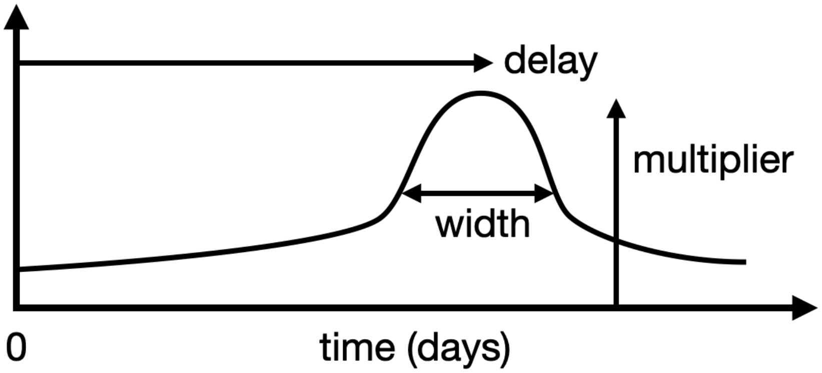


Figure 1: The convolution kernel (defined only for t > 0) provides a time delay, width, and multiplier describing the transition from the contagious pool to the confirmed and the deceased pools with different parameter values.

All transitions can be modeled by the three numbers, as shown in Figure 1:

- Average time delay (μ)
- Variance in time delay (σ^2)
- Overall multiplier (M), i.e., how many people in total transition from one state to the next)

Note that in a usual system of ODEs (as in the SIR/SEIR models), each transition is modeled not by three by just one single parameter, the transition rate. This means that our system of parameters is over-determined compared to a system of ODEs (see below for how we address those concerns using domain knowledge). However, these ODE systems are better-suited to the assumptions of molecular dynamics rather than disease spread over short time scales. For example, a person who becomes contagious has zero chance of being immediately tested and confirmed positive, but this is possible in a rate-based model. Moreover, to account for the shapes of the data that we have, such systems need to postulate at least one intermediate stage between Contagious and Positive or Deceased, rendering them unjustifiably complex, to allow people to transmit from the contagious pool and aggregate in intermediate pools before emitting into the positive and deceased pools.

We describe our model according to Figure 2 based on two observables (confirmed positive cases and deaths) and one non-observable (the number of newly contagious individuals). Because our model is a directed acyclic graph, we can start with a hypothetical profile of the newly Contagious pool and use convolution to determine the Positive and Deceased pools. To obtain the Contagious profile, we integrate the simple ODE below.

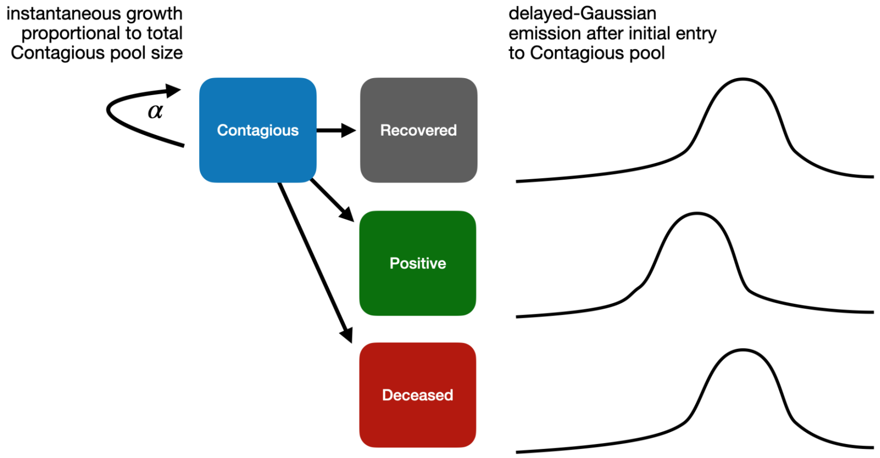


Figure 2: Diagram of the states modeled: people enter the contagious and emit to the Positive and to the Deceased state by delayed-Guassian emission with different delays, widths, and multipliers. However, the Contagious pool itself grows with instantaneous emissions dependent on the total size of the pool. We model emission to the Recovered pool as having the same delay and width as the learned parameters for emission to the Deceased pool, with a multiplier set so all members of the Contagious pool either recover or die. While the Recovered and Deceased pool are mutually exclusive, the Positive pool overlaps with both. Our model only concerns itself with new Positive and Deceased entrants, rather than their cumulative numbers.

$$\frac{dC}{dt} = \alpha(t)C$$

$$\alpha(t) = \frac{\alpha_2 - \alpha_1}{1 + \exp(t_0 - t)} + \alpha_2$$

To define $\alpha(t)$, we use a logistic function in time with fixed width parameter set to 1 day to interpolate between two extremal constant values 1 and 2 for the infectivity rate before and after a Shelter-in- Place order t_0. The difference in the two infection rates indicates the effectiveness of the order on reducing the growth of new infections. The above ODE produces a profile of C(t) that has two growth rates connected using a fixed-width logistic transition between the two growth rates.

The transition of a single individual from the Contagious to Positive and Deceased pools is modeled by a different gaussian kernel for each transition:

$$K(t,t') = \theta(t,t')\frac{M}{\sigma\sqrt{(2\pi}}\exp\left[-\left(\frac{t-t'-\mu}{\sigma}\right)^2\right]$$

where $\theta(t,t')$ is a step function returning 1 for $t' > t$ and 0 otherwise. The kernel is normalized to unit area so that M gives the total number of people who transition from one state to the next. The value provides the average delay between one state and the next, and the value indicates the variance in time delays between one state and the next.

2 Handling Transient Perturbations Around Shelter-in-Place

Our model explicitly does not model the Recovered pool, which means that the initial and final growth rates α_1 and α_2 are effective growth rates that account for emission into the Recovered pool. In fact, the growth rate drawing new people into the Contagious pool (the "leading front") is slightly higher than the effective growth rate, since it must compensate against the loss of individuals to the Recovered pool. In Figure 3, we show an example of how this works: Individuals enter the Contagious pool, then emit either into the Recovered or Deceased pools with a delayed Gaussian probability with values for the delay, width, and multiplier based on fits to the Deceased curve for U.S. totals ($\mu = 19$, $\sigma = 10$, $M = 0.01$). On the other hand, new individuals enter the Contagious pool in proportion to the total size of the pool and this occurs with an instantaneous growth rate without delay.

As shown in Figure 4, the effective growth is demonstrated to be less than the original growth by a fixed amount related to the emission parameters, with an artifact occuring during the change in growth rates that appears as a temporary

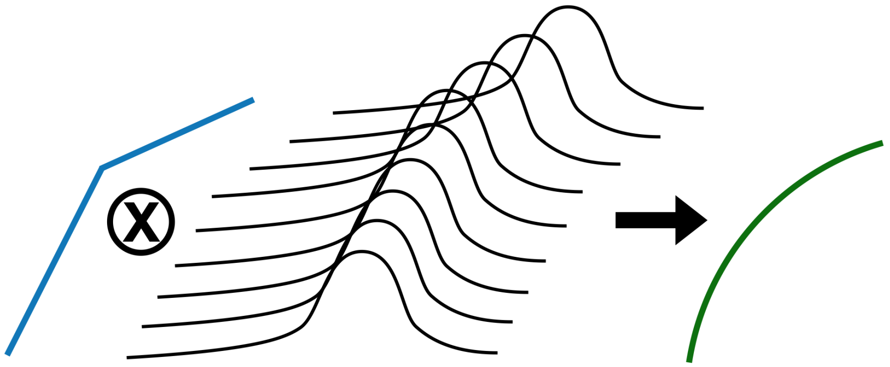


Figure 3: Diagram describing the convolution process. A function with a disjoint derivative (blue, on left), when convolved with a delayed Gaussian kernel (black, middle) produces a smooth curve that will be shifted to the right and be changed in magnitude (green, right).

hump. However, after that hump has been smoothed out by the Gaussian emission into the Positive and Deceased pools, it is no longer visible. The parameters we describe in this document refer to the solid blue curve, which we approximate without the transient artifacts around shelter-in-place. We provide an example solution in Figure 3 against the U.S. totals.

3 Reducing Model Parameters Using Domain Knowledge

The following parameters are determined by the data:

- The slope of positive cases at the beginning and at the end of the shelter-in-place order (after a delay, of course) determine the two growth rates 1 and 2. The value of 2 may not be precisely known until more time has passed to collect more data.
- The change in the slope of positive cases indicates the likely delay between becoming contagious and testing positive. However, if the transition width is sufficiently large, this delay may become ambiguous.
- Similar arguments apply for the parameters describing transitions from the Contagious to the Deceased pools.

In addition, there are also ambiguities in the model since it is overdetermined. In particular, two parameters together describe the relative magnitude of the Contagious and Positive curves. Therefore, we recommend fixing one of these two values:

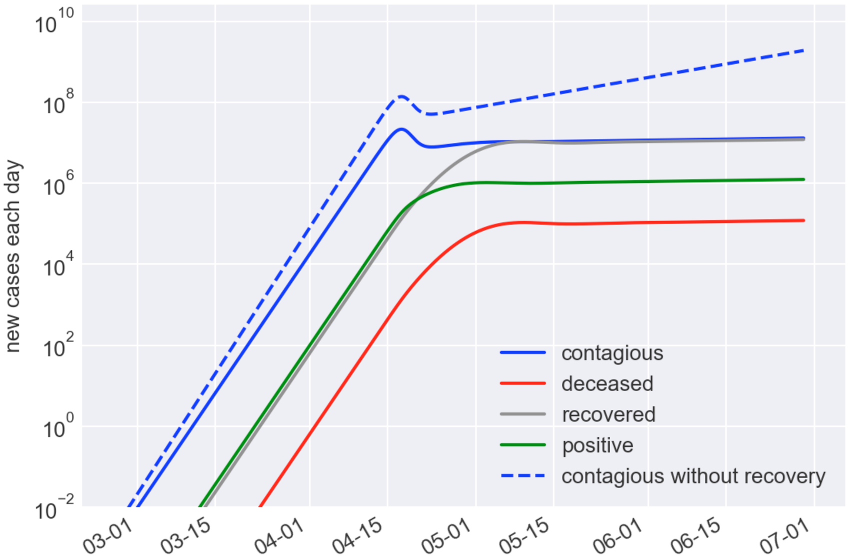


Figure 4: Example solution showing the Contagious, Positive, and Deceased curves for a model incorporating the Recovered emissions. We find that considering the Recovered pool pulls down the "leading front" growth rate into effective growth rates that match the growth rates in our data. Details about emission may or may not produce a visible hump in the unobserved Contagious pool around the time of shelter-in-place, which is then smoothed out through emissions into the Positive and Deceased pools. This is how pools which have reached a flat effective growth rate continue to draw in new individuals, since the raw growth rate for new Contagious individuals is actually slightly higher than what we observe to compensate for loss to the Recovered pool.

- The initial conditions (hypothetical # of newly contagious people) at time (t_0) (Note, this can also be equivalently described as the time in the past that the first individual entered the Contagious pool.)
- The contagious-to-positive multiplier

Since there is no means of knowing the initial conditions, we recommend fixing the contagious-to- positive multiplier to an approximate value supported by evidence, and set the value to 10%. Note that the degeneracy in describing initial conditions (number of newly contagious individuals at a fixed date, or the time in the past that the first individual became contagious) imply that the fraction of contagious individuals who are tested and confirmed necessarily goes down the earlier in the past that the first individual became contagious.

Another ambiguity arises regarding the three parameters:

- The contagious-to-positive delay width
- The contagious-to-positive delay
- The final contagious growth rate

This is because the data supports two groups of solutions:

- Ones with shorter delays and delay widths and with a more-positive final growth rate
- Ones with longer delays and delay widths and with a more-negative final growth rate

which is also true for the contagious-to-deceased transition, where the lower data counts (and therefore greater measurement ambiguity) support the second solution more strongly. However, we can use domain knowledge to identify that since the values supporting the first hypothesis are more-closely aligned with realistic numbers (we doubt that the delay is over a month or that there is over 20 days of variance in the delay).

There are three approaches for handling this ambiguity:

1. Fix some parameters to reduce the expressivity of the model (and increase speed of optimization)
2. Apply priors that rule out undesirable solutions
3. Add terms to the loss function to penalize undesirable solutions

We apply all three techniques (see the next section) and, in particular, fix the contagious-to-positive and contagious-to-deceased emission widths to the approximate values for the U.S. totals: 7 days, or 1 week, for both.

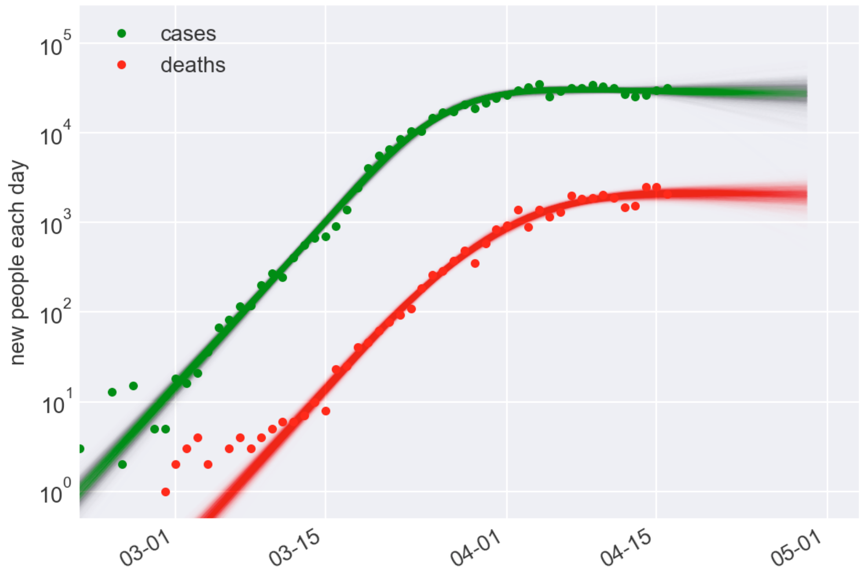
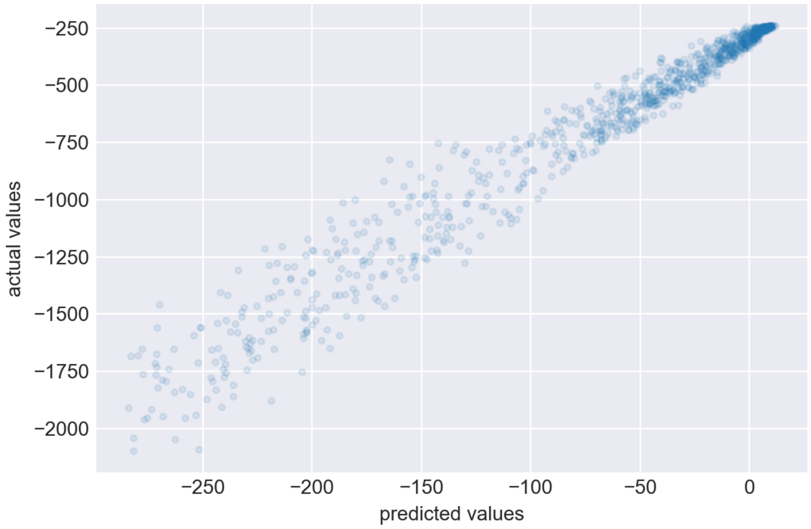


Figure 5: Model Curves and Experimental Data of U.S. COVID-19 Daily Reported Cases and Related Deaths in the U.S.

4 Likelihood Approximation and Incorporating Prior Beliefs

We utilize two approximations to the full likelihood distribution across our parameters, given our observations. The first approximates the likelihood as the distribution of maximum-likelihood-estimates (MLEs) of the parameters on bootstraps of our data, using least-squares error on the difference in the expected and measured log values of confirmed cases and casualties. We consider measurement uncertainty by noting that we expect newly reported cases and deaths to vary up to $\sqrt{N}$ from their underlying values since such counts can be modeled as a Poisson process, whose variance is equal to its measured mean, although the resulting variance ignores other systematic influences on our data such as time-varying testing rates or imperfect tests. Thus, we sample training data with replacement and add samples from a normal distribution with a standard deviation of $\sqrt{N}$ to the recorded values, run our simulation using a discrete ODE solver and convolution library, and minimize the least-squares error over newly confirmed cases and deaths after the 100th case and death has been identified, respectively. We show an example set of solutions for U.S. totals in Figure 5. From the resulting parameter estimates, the bootstrapping technique has been shown to accurately describe the distributions up to second-order statistics (and possibly higher)[5, 4, 9].



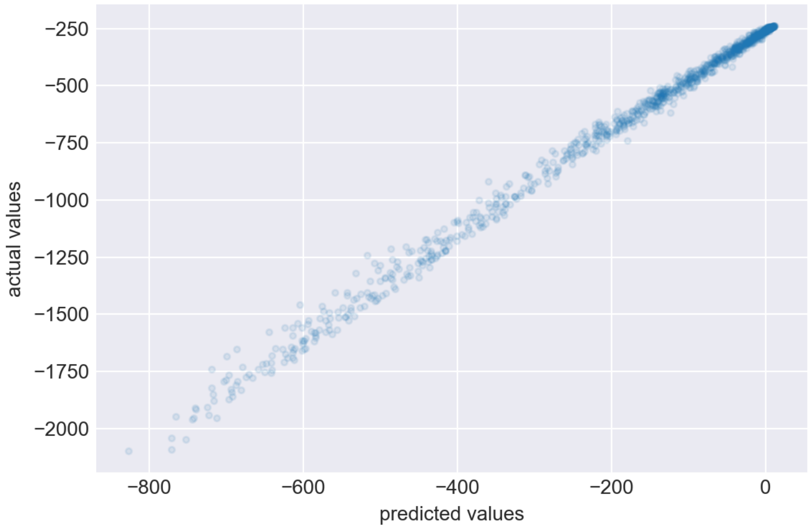


Figure 6: Actual vs. predicted values for the likelihood function for U.S. totals as modeled by a one-component Gaussian mixture model. At top, we show results with a diagonal covariance matrix, and at bottom we show the correlation matrix for the resulting fit.

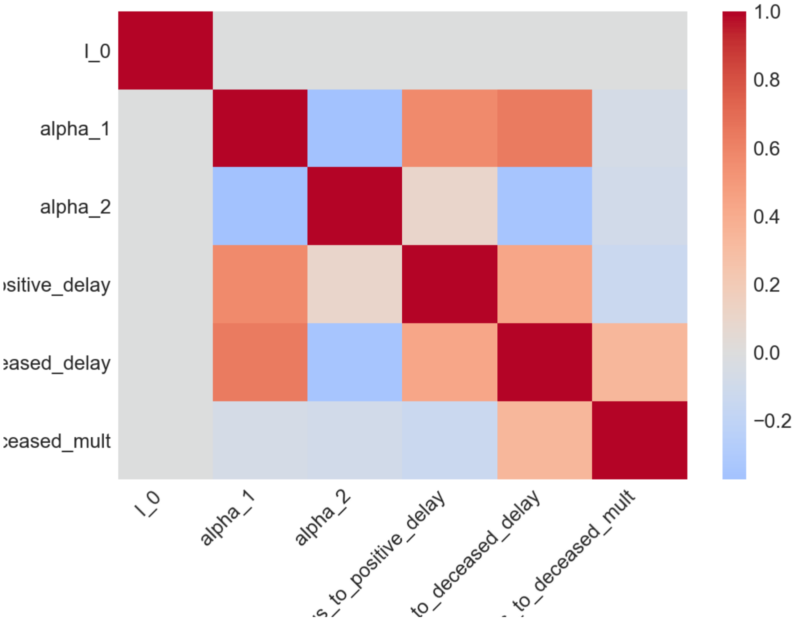


Figure 7: Correlation matrix for model parameters for U.S. totals after being fit to a GMM with a full covariance matrix. We see that α_1 is strongly correlated with the two delays.

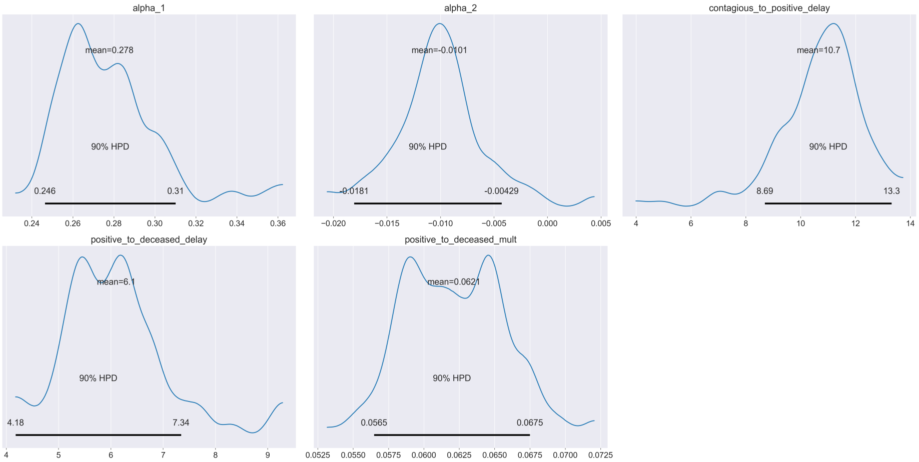


Figure 8: Model Parameter Estimates for U.S. totals after incorporating priors

In the second approximation method, we sample the likelihood function at many points, then resample based on the propensity of those choices and their likelihood, and finally compute aggregate statistics (means and covariance matrix) on the resampled parameters, as well as highest probability density credibility intervals. We also have two methods for generating our samples: the first samples from a normal distribution around the MLE with corresponding propensities coming from the multivariate norm probability density function (PDF), and the second employs the random walk Markov chain Monte Carlo Metropolis-Hastings algorithm[1] with constant propensity, since the resulting distribution should match the underlying probability density directly. This likelihood function is calculated as the product of the PDF of norms centered at each observed data point (N), with standard deviation $\sqrt{N}$, evaluated at the values returned by our simulation. From these samples, we then re-sample based on the likelihood values (normalized to one over all samples) multiplied by their inverse propensities, and run aggregate statistics on the resampled points to create a multivariate norm (MVN) distribution that is representative of our data. In Figure 6, we show the predicted and actual values of the likelihood for U.S. totals and show that the MVN and MCMC approximations demonstrate reasonable accuracy. Moreover, adding the additional parameters by going from a diagonal to a full covariance matrix further improves the fit. In Figure 7, we show the full correlation matrix for the model parameters. We see that α_1 is strongly correlated with the two delays, which will explain deviation from the bootstrap approximation in subsequent analysis.

Both approximation methods contribute hyperparameters to our models: the number of bootstraps (100), the number of likelihood samples (20k), the number of likelihood re-samples (20k), the number of MCMC samples (16k), and the number of burn-in (discarded) MCMC samples (4k).

Once the likelihood distribution has been approximated, according to Bayes theorem,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

incorporating priors on our parameters is as easy as multiplying the two functions element-wise and normalizing a posteriori. In Figure 8, we show the full likelihood approximation (or, equivalently, the posterior with uniform priors), and discuss the physicality of parts of the distributions, which the reader can either incorporate or discount based on their intuition. Updates to the distributions of all parameters based on assumed priors can be obtained by weighted sampling of the bootstraps according to those individual priors on each parameter, since, for example, ruling out overly long delays between becoming contagious and being confirmed positive will also rule out other parameter distributions, such as overly long variances in the delay. We employ the following priors, using uniform distributions (so their shapes don't affect the posterior distribution) with the following bounds:

- $0 < \alpha_1 < 1.0$
- -0.5 $< \alpha_2 <$ 0.5
- 0 $<$ contagious-to-positive delay $<$ 20
- 5 $<$ contagious-to-deceased delay $<$ 40
- 0 $<$ contagious-to-deceased multiplier $<$ 0.1

Note that if we rely on using our priors to filter out undesirable solutions, we can waste a substantial amount of computational resources producing undesirable candidates in our approximations which we only later discard. Worse, it is possible to create unphysical results with our simulations, and these unphysical results can be reached through the optimization methods we employ even when they start with valid parameters. In particular, unphysical results occur when the parameters can produce negative numbers of confirmed cases and deaths, and when the parameters give a later contagious-to-positive than contagious-to-deceased delay. To avoid computing such solutions, we add a term to our loss function which sums the negative value of all predicted values that are below zero, and another term which provides the difference between the contagious-to-positive and contagious-to-deceased delays when the former is larger than the latter. These loss functions have a zero derivative when parameters are valid and push parameters back to the valid range when they are invalid, thus avoiding skewing our results while enforcing physicality. The remaining contributors to the loss function remain the same, however we only contribute terms when the simulation provides positive counts, since the loss function is based on distances from the log results of our simulations, which is undefined for negative values.

5 Results

Examining the parameter estimates in Figures 8, 11, and 12, we see that California shows a much greater response delay to the shelter-in-place order than in New York (March 19th and 20th, respectively) at 12 days compared to 4.2 days (with U.S. totals in the middle at 9.2 days). In addition, the relative ordering of the original growth rates (California = 22.7% < U.S. = 27.8% < New York = 36.9%) reflect the strong population density in New York. However, this ordering is reversed in the final growth rates (New York = +0.435% < California = 1.01% == U.S. = 1.01%) suggesting that New York has dramatically reduced spread.

All three datasets show a similar relative delay between being a case being positively confirmed and a resulting death at approximately one week (7-8 days). When looking at the relative multiplier between the positive and deceased numbers, which acts as an analog to the case fatality rate, we find that California's estimates are much lower (4.7% for California vs. 6.21% for the U.S. and 6.45% for New York). If tests are limited and reserved only for the most severely ill, we would expect the relative multiplier to increase, suggesting that New York and the U.S. may be capping testing more severely than California.

In Figures 9 and 10 we provide results for all 50 states, Washington D.C., and U.S. totals for the final growth rate and the relative multiplier between the positive and deceased pools, and for both the bootstrapping and MCMC approximations. We find agreement between the two approximations in most cases, and see a wide disparity among the different regions, which the reader is encouraged to interpret.

References

[1] Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57(1):97–109, April 1970.

[2] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control*. Dynamics and Control. OUP Oxford, 1992.

[3] Robert Martinez Douglas Mason. Bayesian ode/convolution models for estimating underlying growth of covid-19 and its uncertainty. *https://github.com/douglasmason/covid_ model.*

[4] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[5] Bradley Efron. Second thoughts on the bootstrap. *Statist. Sci.*, 18(2):135–140, 05 2003.

[6] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.

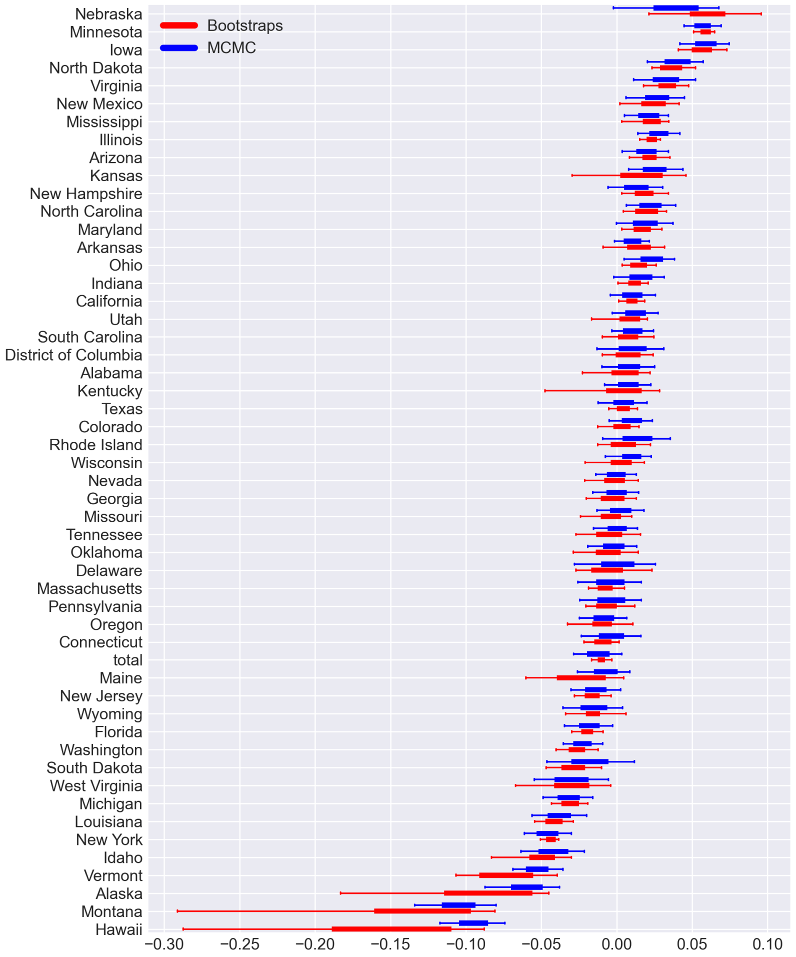


Figure 9: Model parameter estimates for α_2 (the current growth rate of COVID-19) for each of 50 U.S. states, Washington D.C., and U.S. totals with 5%, 25%, 50%, 75%, and 95% percentiles, ranked from highest to lowest median, and shown with both the bootstrap and the MCMC approximations. We find that both approximation methods agree with each other. We see strongest growth in Nebraska, Minnesota, and Iowa, and lowest growth in Alaska, Montana, and Hawaii.

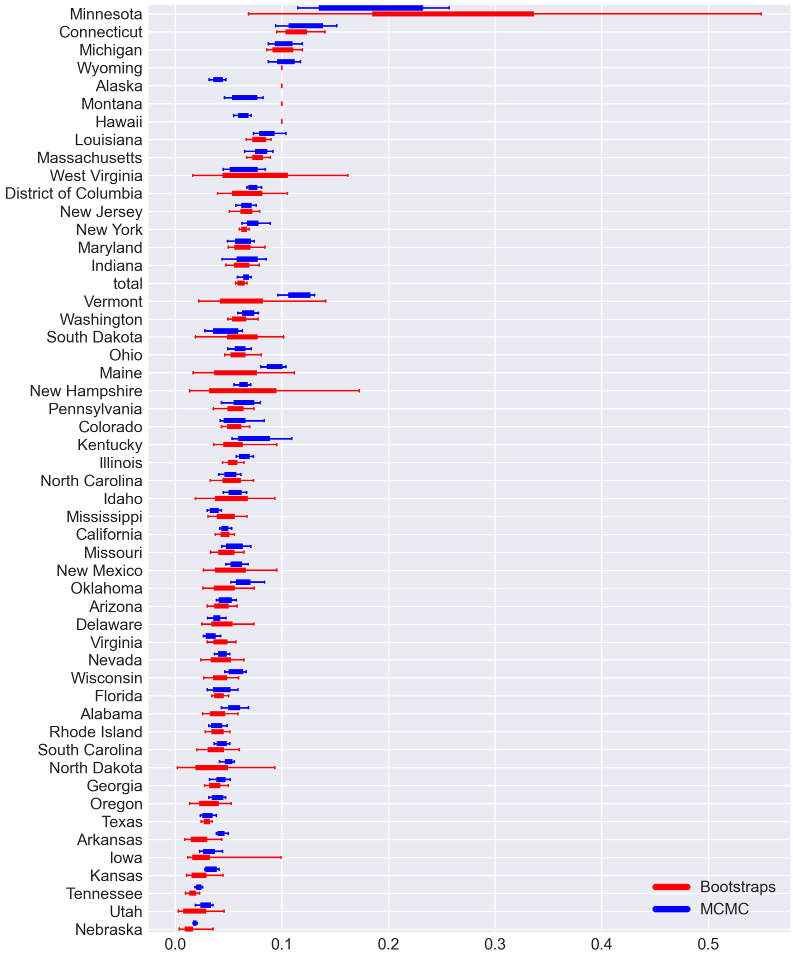


Figure 10: Model parameter estimates for the positive-to-deceased multiplier, which is an analog to the case fatality rate and an indicator of testing restrictions when the value is high. We provide results for each of 50 U.S. states, Washington D.C., and U.S. totals with 5%, 25%, 50%, 75%, and 95% percentiles, ranked from highest to lowest median, and shown with both the bootstrap and the MCMC approximations. We find that the bootstrap method gives us larger variances, and fails to find variance for states with very low death counts (Wyoming, Alaska, Montana, and Hawaii) and only returns the initial value (10%). Overall, we see a strong skew towards East Coast states with higher estimations.

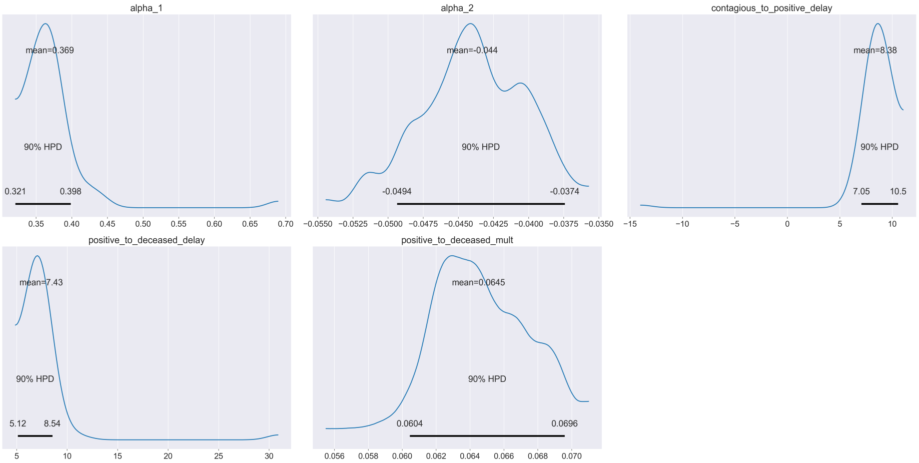


Figure 11: Model Parameter Estimates for New York after incorporating priors

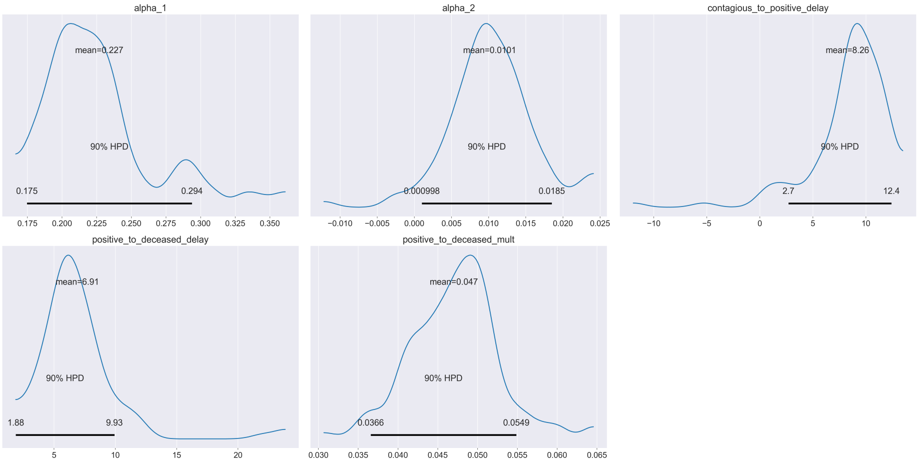


Figure 12: Model Parameter Estimates for California after incorporating priors

[7] Eduardo Massad. An introduction to infectious diseases modelling. by e. vynnycky and r. white (400 pp.; Â£29.95; isbn 978-0-19-856576-5 pb). oxford: Oxford university press, 2010. *Epidemiology and Infection*, 139(7):1126–1126, 2011.

[8] Sarah Almukhtar Keith Collins Danielle Ivory Mitch Smith, Karen Yourish and Amy Harmon. Coronavirus (covid-19) data in the united states. *https://github.com/nytimes/covid-19-data*.

[9] Donald B. Rubin. The bayesian bootstrap. *Ann. Statist.*, 9(1):130–134, 01 1981.