Models and Optimization for Forecasting COVID-19

Andrei Graur, **Arec Jamgochian**, **Bernard Lange** and **Kyriakos Lotidis**
School of Engineering, Stanford University
{agraur, arec, blange, klotidis}@stanford.edu

Abstract

COVID-19 pandemic has highlighted the importance of prediction algorithms in modeling the future trajectory of infectious disease, namely predicting future cases, hospitalizations, and deaths. Such knowledge can be used to effectively and efficiently manage the government's response planning. We propose a Gamma Model which captures the true infection rate of the infectious disease and demonstrate its performance on COVID-19 compared with L1-regularized autoregressive moving average model and a SIRD model. We provide our implementation at https://github.com/jamgochiana/CovidModeling.

1 Introduction

One year of the COVID-19 pandemic has led to more than 3 million deaths and 150 million cases worldwide. It has rapidly upended people's lives and transformed the world in an unprecedented way. The exponential rate of increase of new infections, combined with the significant percentage of patients who require hospitalization and respiratory support, has exposed the weaknesses of pandemic preparedness and response planning.

To protect public health, the governments have imposed strict measures to reduce social contact and control the spread of the virus. Undoubtedly, the severity and the timing of such restrictions should be motivated by the true state of the pandemic, namely, the accurate number of COVID-19 cases. However, due to the large percentage of asymptomatic cases and limited testing capabilities, such information is very challenging to acquire.

As part of this project, we aim to obtain a more accurate estimate of the true infection rate of COVID-19, which would hopefully enable a more proactive, efficient, and effective introduction of such restrictions.

2 Background

Since the outbreak of COVID-19, many models for forecasting and for studying the underlying dynamics and characteristics of the disease have been proposed. These models can be separated into two main categories: (1) SIR-based models, and (2) curve-fitting models.

In order to study the large-scale epidemiological aspects of the pandemic, it is convenient to use mean field analysis models with Markovian structure. The so-called SIR model is perhaps the most commonly used one because, while relatively simple, it captures the main drivers of the macroscopic spreading process. This model, introduced by Kermack et al., studies the dynamics of the density of the susceptible population, denoted by S, the density of the infected population I, and the remaining population which is either recovered or deceased, denoted by R. This is a very general model that can capture the dynamics of most infectious deceases, and many variants have been proposed to model COVID-19.

Chowdhury et al. implement a SEIR (Susceptible-Exposed-Infected-Recovered) compartmental model to simulate the ICU demand and deaths in different non-pharmacological interventions (NPI) scenarios in 16 countries from diverse regions and economic status. In a different line of work, the modified SIRD model, that takes into account the *deaths* associated with the virus, is used to analyze time series data and predict the total cases, deaths and the recovered population (Sen and Sen, 2021). Zou et al. augments the SEIR model to capture the *undetected* infections, proposing the SuEIR model. Using machine learning algorithms, they try to predict the future deaths and peak dates of active cases across the United States. There are many other modifications of this dynamical model, where the individuals can be in more states (e.g. quarantine or hospitalization) to effectively adapt to the special properties of COVID-19 (Cordelli et al., 2020; Gribaudo et al., 2021; Mandal et al., 2020). Finally, a new axiomatic model of epidemic development HIT captures special characteristics of COVID-19, like delayed detection and asymptomatic virus holders (Nesterov, 2020).

Data-fitting models are the second major class used for forecasting COVID-19 cases. Hoseinpour Dehkordi et al. use linear regression to project future cases and deaths, collecting data from many different countries. Weinberger et al. use Poisson regression models to estimate excess deaths associated with COVID-19 in the United States. Following a similar approach, autoregressive moving average (ARMA) models and some variants have been widely used to forecast the daily global incidence of COVID-19 (Yousaf et al., 2020; Alzahrani et al., 2020; Ceylan, 2020; Pourghasemi et al., 2020).

3 Methodology

In this section we describe the three models used to fit different models and make predictions on subsets of new case, total hospitalization, and/or new death data. We formu-

late a convex, L1-regularized autoregressive moving average (ARMA) model, our own convex model which takes into account lagged dynamics, and a non-convex SIRD model. In all our models, we denote the number of new cases, total hospitalizations, and new deaths at time t as c_t, h_t, and d_t respectively.

3.1 Regularized Autoregressive Moving Average

An ARMA model is a function which linearly regresses features from L previous time steps to features at the next time step. For the purposes of comparing to our own lagged-dynamics model presented in Section 3.2, we use L steps of hospitalizations and deaths to regress to next step hospitalizations and deaths. We can describe this predictor function $\mathbf{f}_{\mathbf{\Theta}} : \mathbb{R}^{2L} \to \mathbb{R}^2$ as follows:

$$\begin{bmatrix} h_{t+1} \\ d_{t+1} \end{bmatrix} = \mathbf{f}_{\mathbf{\Theta}}(\mathbf{h}_{t-L+1:t}, \mathbf{d}_{t-L+1:t}) = \mathbf{\Theta}^{\intercal} \begin{bmatrix} \mathbf{h}_{t-L+1:t} \\ \mathbf{d}_{t-L+1:t} \\ 1 \end{bmatrix}, \tag{1}$$

where $\mathbf{\Theta} \in \mathbb{R}^{2L+1 \times 2}$ is a matrix encoding learned model weights.

To learn model weights from data, we optimize an L1-regularized least-squares objective (Lasso). The L1-regularization is convenient for enforcing sparsity in our weights and preventing overfitting. If we form a dataset ($\mathbf{X} \in \mathbb{R}^{n \times 2L}, \mathbf{Y} \in \mathbb{R}^{n \times 2}$) of all n L-step inputs and 1-step outputs of hospitalizations and deaths in a particular time-series, we may fit $\mathbf{\Theta}$ by minimizing the objective

$$\min_{\mathbf{\Theta}} \frac{1}{n} \|[\mathbf{X}\ \mathbf{1}]\mathbf{\Theta} - \mathbf{Y}\|_2^2 + \lambda \|\mathbf{\Theta}\|_1, \tag{2}$$

where λ is a tunable relative weighting parameter.

This optimization objective can be solved efficiently with coordinate descent.

3.2 Gamma model

The core idea of our Gamma model is to first split the population into two types: the vulnerable population, V, and the non-vulnerable population, U. We assume that every individual in V, upon infection, will end up with a hospitalization/severe case, in a matter of Δ_1 days (a model constant we discuss below) and that a record will be kept about their hospital presence. For the U-type individuals, we assume no hospitalization would follow. It is natural to have these two types of individuals in our model, as regarding COVID-19, we often think of the older people with pre-existing conditions as being a lot more susceptible to severe illness, and the young people without pre-existing conditions as the ones who are a lot more sheltered from grave consequences.

Next, we present some of the notation used in the model. We split our quantities to three parts: (1) the observable quantities, (2) the non-observable variables, (3) the constants. Note that the observable quantities are extracted directly from the data, while the non-observable ones capture the underlying structure of out model and are computed through the relations described in the experiments section.

- Observable Quantities
  - d_t - number of new deaths observed during day t
  - h_t - number of people currently in the hospital during day t
- Non-Observable Variables
  - T_t - number of infections (past and present) that happened up to day t
  - V_t - number of type V (vulnerable) individuals that are actively infected during day t
  - U_t - number of type U individuals that are actively infected during day t
- Constants
  - $\Delta_1 = 11$ - delay in detecting a severe case since the occurrence of the infection
  - $\Delta_2 = 18$ - time it takes for a death to occur since the infection
  - $\Delta_3 = 14$ - how long an infection will last
  - $p = 0.02$ - probability of death given infection

The values of the constants above were set by taking into consideration estimates (obtained by healthcare professionals through extensive empirical evidence) of the following: average number of days it takes to require hospital care if infected, average number of days it takes for a death to occur, average number of days an infected person is contagious, and probability of dying upon infection.

Assumptions

Our model makes the following assumptions:

1. $h_t = V_{t-\Delta_1}$. This signifies that every case among type V individuals is detected in Δ_1 days. Intuitively, this makes sense, as we assume each type V individual will eventually develop a severe case, which will be detected when they arrive at the hospital.
2. $d_t = p(T_{t-\Delta_2} - T_{t-\Delta_2-1})$. This assumption is equivalent to saying that given that a death occurs for a certain individual, it is detected in exactly Δ_2 days.
3. There exist, for each day t a matrix $\mathbf{\Gamma}_t$ such that

$$\begin{bmatrix} V_{t+1} \\ U_{t+1} \end{bmatrix} = \mathbf{\Gamma}_t \begin{bmatrix} V_t \\ U_t \end{bmatrix} + \epsilon_t,$$

$$\epsilon_t \sim \mathcal{N}(0, \sigma^2 \begin{bmatrix} V_t & 0 \\ 0 & U_t \end{bmatrix})$$

   Note that $\mathbf{\Gamma}_t = \begin{bmatrix} \gamma_{t,11} & \gamma_{t,12} \\ \gamma_{t,21} & \gamma_{t,22} \end{bmatrix}$ is a time-dependent matrix.
4. The entries of $\mathbf{\Gamma}_t$ do not change too much over a time interval of L days (for some fixed L). We expand on this in the experiments section.

Assumption 3 refers to how we model the dynamics of the infection evolution. One of our core objectives here is to estimate these gamma parameters which determine the dynamics of how the infection evolves. This can help with getting a better estimate of the true current case counts as well as predicting how the infection will grow over time. Assumption 4 refers to the fact that the underlying dynamics of the infection cannot change too drastically over a short time period of L days. In the experiments section, we describe how we use this assumption to compute the quantities $\mathbf{\Gamma}_t$.

Alternative Gamma model

We also consider a simpler model, where we only have two dynamic-related parameters for each day, $\mathbf{\Gamma}_t = [\gamma_{t,1}, \gamma_{t,2}]^\intercal$. We model

$$\begin{bmatrix} V_{t+1} \\ U_{t+1} \end{bmatrix} = (V_t + U_t)\mathbf{\Gamma}_t + \epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2(V_t + U_t)I)$.

The rest of the model, namely constants, and the split in the U, V types, is the same as in the original Gamma model. Note that this is a more simplistic version of the Gamma model, because here we assume the pattern of interactions of an individual with members of their type is the same as for interactions with members from the opposite type. In this model, $\gamma_{t,1}$ describes how the infection among the V-type will evolve based on the total number of active infections at day t'. By contrast, in the Gamma model, the value of $V_{t'+1}$ depends on both $U_{t'}, V_{t'}$, as we consider the cross-interactions $U - V$ to be different than the interactions $V - V$.

Comparing the alternative Gamma Model with the original is valuable, as it helps us understand if the nature of interactions $V - V$ and $U - U$ are fundamentally different than $U - V$. Intuitively, we expect the answer to be in the affirmative, as we would expect the type V individuals to shelter more from type U individuals, who are more active and thus more susceptible to be infected, than from individuals in V, who we expect to be more cautious about getting infected.

3.3 SIRD model

A Susceptible-Infectious-Recovered-Deceased (SIRD) model is a deterministic compartmental model of the infectious disease where the population is divided into the following compartments: susceptible $\boldsymbol{S(t)}$, infectious $\boldsymbol{I(t)}$, recovered $\boldsymbol{R(t)}$, and deceased $\boldsymbol{D(t)}$. An extension of the SIRD model also contains confirmed cases $\boldsymbol{C(t)}$. Each of the aforementioned variables accounts for the population number at a specific point in time. The SIRD model is defined with the following partially observable nonlinear system:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta IS}{N} \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I - \mu I \\ \frac{dR}{dt} &= \gamma I \\ \frac{dD}{dt} &= \mu I \\ \frac{dC}{dt} &= \frac{\beta IS}{N} \end{aligned} \tag{3}$$

where β, γ, and μ are constants that we must optimized for, and only C and D are observed. In our project, we discretize the dynamics with an integration scheme (e.g. 4th order Runge-Kutta) and end up with the generalized discrete-time state-space system

$$\begin{aligned} \mathbf{x}_{t+1} &= f_\theta(\mathbf{x}_t, t) + w_t \\ \mathbf{y}_t &= g_\theta(\mathbf{x}_t, t) + v_t, \end{aligned} \tag{4}$$

where $\mathbf{x}_t \in \mathbb{R}^n$ is the full state at time t (i.e. $[S_t, I_t, R_t, c_t, d_t]^\intercal$) and $\mathbf{y}_t \in \mathbb{R}^m$ is the observation at time t (i.e. $[c_t, d_t]^\intercal$). f_θ can be defined with discretization of the dynamics in Equation (3), θ contains model parameters (β, γ, and μ), and w_t and v_t are process and observation noises, respectively.

4 Experiments

In our experiments, we compare the three models described in Section 3 based on how they perform at the task of forecasting true unseen COVID-19 statistics given past statistics.

4.1 Dataset

To train and test our models, we use the *Our World in Data* Covid-19 dataset (Hasell et al., 2020)[1]. This dataset consists of many statistics reported daily for a number of countries through the course of the pandemic. From this dataset, we extract time-series of new cases, total hospitalizations, and new deaths reported daily in the United States, the United Kingdom, and Italy.

4.2 Metrics

To compare methods, from each country location, we randomly select $B = 25$ sets of 42-day input series to use for training, and the corresponding $K = 7$ following days to use for testing. We train model parameters over each input sequence and propagate our learned model forward to make a prediction of the COVID-19 trajectory over the next 7 days. We then report the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE), which quantify linear- and quadratic-like errors. These metrics are defined as follows:

$$MAE = \frac{1}{BK}\sum_{i=1}^{B}\sum_{\tau=1}^{K} \mid y_{t+\tau}^i - z_{t+\tau}^i \mid, \tag{5}$$

$$MAPE = \frac{1}{BK}\sum_{i=1}^{B}\sum_{\tau=1}^{K} \mid \frac{y_{t+\tau}^i - z_{t+\tau}^i}{y_{t+\tau}^i} \mid, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{BK}\sum_{i=1}^{B}\sum_{\tau=1}^{K} \left(y_{t+\tau}^i - z_{t+\tau}^i\right)^2}, \tag{7}$$

where $y_{t+\tau}^i$ indicates the true value of a particular feature at time $t+\tau$ in the i-th problem, and $z_{t+\tau}^i$ indicates the forecast at that same index in time.

4.3 Model details

ARMA

To fit the ARMA model, we apply a rolling window within each 42-day training time-series to form a dataset consisting $L = 14$-day input and 1-day output pairs. We then fit model parameters using the Scikit-learn implementation of Lasso (Pedregosa et al., 2011), which optimizes efficiently using coordinate descent. We use a regularization parameter $\lambda = 5$. To make a prediction, we use the final 14 days of training data to input to our learned model, and iteratively propagate our learned model forward 7 days by appending new outputs to the model input.

[1] Available at https://covid.ourworldindata.org/

Gamma

To fit the Gamma model, we first reconstruct the values of $(U_{t'}, V_{t'})$ up to day $t - \Delta_2$. At day t, we have seen $h_{t'}$ and $d_{t'}$ for all $t' \leq t$ and, therefore, from $h_{t'}$'s, we recover $V_{t'-\Delta_1}$ for $t' \leq t$. Note that $\Delta_2 > \Delta_1$, so at day t, we know $V_{t-\Delta_1}$ but also $V_{t-\Delta_2}$. With the data received so far, we aim to compute $U_{t'-\Delta_2}$ for all $t' \leq t$. To do so, we first compute $U_{t-\Delta_2} + V_{t-\Delta_2}$. This will be enough to compute $U_{t-\Delta_2}$ (and the earlier $U_{t'}$'s) because, as we have pointed out, we know the value of $V_{t-\Delta_2}$. To compute the sum $U_{t-\Delta_2} + V_{t-\Delta_2}$, we use the data regarding deaths to compute the number of people who were infected at most Δ_3 days prior to day $t - \Delta_2$. Thus, the quantity we want to estimate is $\sum_{t'=t-\Delta_2-\Delta_3}^{t-\Delta_2}[T_{t'} - T_{t'-1}]$ (new infections for each day $t' \in \{t - \Delta_2 - \Delta_3, t - \Delta_2 - \Delta_3 + 1, ..., t - \Delta_2\}$. To do so, from our second modeling assumption, we have that

$$\sum_{t'=t-\Delta_2-\Delta_3}^{t-\Delta_2}[T_{t'} - T_{t'-1}] = \sum_{t'=t-\Delta_3}^{t}\frac{1}{p}d_{t'}$$

Hence, we estimate

$$U_{t-\Delta_2} = \sum_{t'=t-\Delta_3}^{t}\frac{1}{p}d_{t'} - h_{t-\Delta_2+\Delta_1}$$

Again, $h_{t-\Delta_2+\Delta_1} = V_{t-\Delta_2}$ as we have access to this at day d because $\Delta_2 > \Delta_1$. After reconstructing the values of $(U_{t'}, V_{t'})$ up to day $t - \Delta_2$, we focus on the last $L = 7$ days for which the values have been completed, namely $V_{t-\Delta_2-L:t-\Delta_2}$ and $U_{t-\Delta_2-L:t-\Delta_2}$. We then formulate two optimization problems to determine the underlying dynamics of the infection in this time-span.

There are two ways we are going to estimate the values $\mathbf{\Gamma}_{t'}$ for $t' = t-\Delta_2-L, t-\Delta_2-L+1, ..., t-\Delta_2$. What these two have in common is the intuition that for a certain number of days, these values should not change too much, meaning that each entry in $\mathbf{\Gamma}$ is almost constant in any time-span of L days. The two formulations we will be using are the following:

1. $$\min_{\mathbf{\Gamma}} \sum_{t'=t-\Delta_2-L}^{t-\Delta_2-1} \| \begin{bmatrix} V_{t'+1} \\ U_{t'+1} \end{bmatrix} - \mathbf{\Gamma} \begin{bmatrix} V_{t'} \\ U_{t'} \end{bmatrix} \|_2^2$$

   This formulation assumes $\mathbf{\Gamma}_{t'}$ is a constant matrix for $t' = t - \Delta_2 - L, t - \Delta_2 - L + 1, ..., t - \Delta_2$

2. $$\min_{\mathbf{\Gamma}} \sum_{t'=t-\Delta_2-L}^{t-\Delta_2-1} \| \begin{bmatrix} V_{t'+1} \\ U_{t'+1} \end{bmatrix} - \mathbf{\Gamma}_{t'} \begin{bmatrix} V_{t'} \\ U_{t'} \end{bmatrix} \|_2^2 + \lambda \sum_{t'=t-\Delta_2-L}^{t-\Delta_2-1} \|\mathbf{\Gamma}_{t'} - \mathbf{\Gamma}_{t'+1}\|_1$$

   Here, we add an L1 term penalizing day-to-day change in $\mathbf{\Gamma}_t$ to enforce smoothness, with weight λ.

In terms of computing the optimal values for Γ, for both optimization tasks, we employ the gradient descent method (or subgradient descent for the case of the second formulation). We set the number of iterations to be 10^5, and we set our stepsize to be $\alpha_k = \frac{\alpha}{\sqrt{k}}$ for $\alpha = 0.1$. The reason we use this method is that CVXPY was running slowly and would not find an accurate enough solution in the default number of iterations, and increasing the iterations even more was slowing down the program even more.

To predict forward the values of $\hat{V}_{t'}, \hat{U}_{t'}$ for $t' > t - \Delta_2$, we assume the gamma values are constant and simulate the evolution of the infection using the computed value of $\mathbf{\Gamma}_{t-\Delta_2}$. More specifically, we predict

$$\begin{bmatrix} \hat{V}_{t-\Delta_2+k} \\ \hat{U}_{t-\Delta_2+k} \end{bmatrix} = \mathbf{\Gamma}_{t-\Delta_2}^k \begin{bmatrix} V_{t-\Delta_2} \\ U_{t-\Delta_2} \end{bmatrix}.$$

An important aspect to note is that, as we have mentioned before, we know the ground truth of values $V_{t-\Delta_2+1:t-\Delta_1}$. Yet, for the prediction phase, we pretend we do not have access to these, predict $\hat{V}_{t-\Delta_2+1:t-\Delta_1}$ as described above, and then proceed to predict $\hat{V}_{t-\Delta_1+1:t-\Delta_1+K}$, so for $K = 7$ days ahead of time t. As far as deaths are concerned, our model is not equipped to estimate very accurately daily new deaths, so for day $d > t$, we predict the number of deaths to be $p \cdot (\hat{V}_{d-\Delta_2} + \hat{U}_{d-\Delta_2})/14$, so as to account for the fact that $\hat{V}_{d-\Delta_2}, \hat{U}_{d-\Delta_2}$ account for active infections (time span of 14 days).

SIRD

To fit the SIRD model, we use Certainty-Equivalent Expectation Maximization (CE-EM) (Menda et al., 2020). CE-EM iteratively perform a two-step procedure—the E-step holds θ constants and infers the unobserved state variable $\mathbf{x}$, while the M-step optimizes for θ. CE-EM differs from vanilla Expectation Maximization (EM) in that it assumes that the distribution of the states conditioned on observation is a Dirac delta function. That is, CE-EM only maintains the most likely estimate for $\mathbf{x}$, making EM much more tractable.

CE-EM optimizes the following objective function with block coordinate ascent:

$$\begin{aligned} J(\mathbf{x}_{1:t}, \theta) = \log p(\mathbf{x}_1) + \sum_{t=1}^{T} \log p_v(\mathbf{y}_t - g_\theta(\mathbf{x}_t, t)) \\ + \sum_{t=1}^{T-1} p_w(\mathbf{x}_{t+1} - f_\theta(\mathbf{x}_t, t)) \end{aligned} \tag{8}$$

Once model parameters θ and most likely state values $\mathbf{x}_{1:t}$ are learned, we can propagate the dynamics and observation functions forward to form a forecast. That is, we can form a 7-day prediction by applying f_θ and g_θ sequentially for 7 days from $\mathbf{x}_t$ (without injecting any noise).

4.4 Results and Discussion

To compare the forecast accuracy of the different models learned, we report the mean of metrics calculated across $B = 25$ series in three countries in Table 1. Since all models forecast deaths, we use death prediction as a benchmark in our discussions. In Fig. 1, we visualize a single 7-day death and hospitalizations predictions for all models.

With the United States time-series, we notice that the SIRD model fit with CE-EM performs better across all metrics, with the Gamma models performing the next best. With the United Kingdom time-series, while the SIRD model again performs the best, the ARMA model is much more comparable to the Gamma models. Finally, with the Italy time-series, the ARMA model outperforms the other two. The results indicate that no model uniformly outperforms the others. Furthermore, we notice by examining the MAPE that at best, we

Table 1: Metrics for 7-day forecast performance computed for different methods in the United States, the United Kingdom, and Italy. Means and standard deviations are computed from $B = 25$ trials.

		Cases			Hospitalizations			Deaths		
	Model	**MAE**($\times 10^5$)	**MAPE**	**RMSE**($\times 10^5$)	**MAE**($\times 10^3$)	**MAPE**	**RMSE**($\times 10^3$)	**MAE**	**MAPE**	**RMSE**
USA	ARMA	—	—	—	$1.714_{\pm 0.284}$	$0.029_{\pm 0.003}$	$2.033_{\pm 0.342}$	$262.831_{\pm 59.965}$	$0.339_{\pm 0.165}$	$331.926_{\pm 72.708}$
	Gamma4	—	—	—	$5.963_{\pm 1.333}$	$0.081_{\pm 0.010}$	$6.191_{\pm 1.375}$	$154.294_{\pm 25.775}$	$0.102_{\pm 0.014}$	$168.289_{\pm 25.841}$
	GammaL1	—	—	—	$4.871_{\pm 0.895}$	$0.077_{\pm 0.008}$	$5.086_{\pm 0.930}$	$142.166_{\pm 21.291}$	$0.099_{\pm 0.013}$	$158.420_{\pm 21.782}$
	Gamma2	—	—	—	$9.191_{\pm 2.145}$	$0.148_{\pm 0.021}$	$9.398_{\pm 2.187}$	$151.172_{\pm 24.563}$	$0.102_{\pm 0.013}$	$162.045_{\pm 24.967}$
	SIRD	$1.002_{\pm 0.2252}$	$0.079_{\pm 0.009}$	$1.085_{\pm 0.241}$	—	—	—	$\mathbf{134.874}_{\pm 26.053}$	$\mathbf{0.083}_{\pm 0.011}$	$\mathbf{143.258}_{\pm 26.545}$
UK	ARMA	—	—	—	$0.344_{\pm 0.070}$	$0.042_{\pm 0.006}$	$0.395_{\pm 0.082}$	$35.294_{\pm 15.214}$	$0.422_{\pm 0.262}$	$45.16_{\pm 20.475}$
	Gamma4	—	—	—	$1.883_{\pm 0.586}$	$0.120_{\pm 0.013}$	$1.972_{\pm 0.612}$	$49.746_{\pm 14.038}$	$0.205_{\pm 0.024}$	$51.444_{\pm 14.169}$
	GammaL1	—	—	—	$1.573_{\pm 0.520}$	$0.101_{\pm 0.011}$	$1.664_{\pm 0.549}$	$46.753_{\pm 12.294}$	$0.199_{\pm 0.022}$	$48.600_{\pm 12.501}$
	Gamma2	—	—	—	$2.554_{\pm 0.805}$	$0.250_{\pm 0.033}$	$2.636_{\pm 0.831}$	$41.594_{\pm 12.580}$	$0.173_{\pm 0.021}$	$43.111_{\pm 12.650}$
	SIRD	$0.310_{\pm 0.125}$	$0.221_{\pm 0.027}$	$0.333_{\pm 0.135}$	—	—	—	$\mathbf{27.514}_{\pm 7.714}$	$\mathbf{0.162}_{\pm 0.023}$	$\mathbf{29.482}_{\pm 8.106}$
Italy	ARMA	—	—	—	$0.436_{\pm 0.090}$	$0.100_{\pm 0.030}$	$0.496_{\pm 0.098}$	$\mathbf{14.937}_{\pm 2.692}$	$\mathbf{0.152}_{\pm 0.033}$	$\mathbf{17.420}_{\pm 3.294}$
	Gamma4	—	—	—	$1.571_{\pm 0.403}$	$0.216_{\pm 0.055}$	$1.676_{\pm 0.431}$	$36.240_{\pm 7.100}$	$0.244_{\pm 0.049}$	$37.227_{\pm 7.175}$
	GammaL1	—	—	—	$1.256_{\pm 0.379}$	$0.158_{\pm 0.035}$	$1.334_{\pm 0.410}$	$35.300_{\pm 6.915}$	$0.248_{\pm 0.035}$	$36.447_{\pm 6.991}$
	Gamma2	—	—	—	$3.117_{\pm 0.897}$	$0.296_{\pm 0.049}$	$3.116_{\pm 0.897}$	$32.997_{\pm 6.254}$	$0.235_{\pm 0.048}$	$34.262_{\pm 6.365}$
	SIRD	$0.137_{\pm 0.032}$	$0.144_{\pm 0.018}$	$0.153_{\pm 0.035}$	—	—	—	$28.108_{\pm 6.680}$	$\mathbf{0.151}_{\pm 0.022}$	$30.215_{\pm 7.060}$

can only expect our models to achieve $8 - 15\%$ accuracy in predicting deaths over 7 days. This validates the notion that predicting COVID is hard.

We also observe that there is a large gap in model fitting times between the three models. In Table 2 we report the times taken to fit all $B = 25$ models for each method with the United Sxtates time-series. We notice that the Gamma model can fit 25 time-series about 13 times faster than the ARMA model, and almost 30 times faster than the non-convex SIRD model.

4.5 Strengths and Weaknesses of Gamma Model

First, the strength of the Gamma Model is that, without using any data regarding positive cases, it estimates the number of true active cases on a given day. During day t, our model either has an estimate obtained from data preprocessing of true number of active cases at day d, provided that $d \leq t - \Delta_2$ (our estimate being $U_d + V_d$), or it has an estimate obtained via prediction, namely $\hat{U}_d + \hat{V}_d$. Hence, our model holds an estimate of a key unobserved quantity and makes predictions regarding deaths and hospitalizations based on that. By comparison, the ARMA and SIRD models do not attempt to estimate the true number of active infections in order to make predictions.

Another strength of the Gamma Model is that it attempts to study the dependence of $V_{t'+1}$ on $U_{t'}$ and $V_{t'}$. In other words, the values of $\gamma_{t,11}, \gamma_{t,12}]$ that we compute explain how the infection spreads during the next day both within the population segment V, and from population segment U to segment V. By comparing the results from the 4-parameter Gamma Model experiments (GammaL1 and Gamma4) with the ones from the 2-parameter Gamma Model (Gamma2), we notice that the 4-parameter ones do much better in terms of predicting hospitalizations. This is consistent with our expectations as we postulated the spread of infection from $U \to V$ might not be the same as the one from $V \to V$ or from $U \to U$.

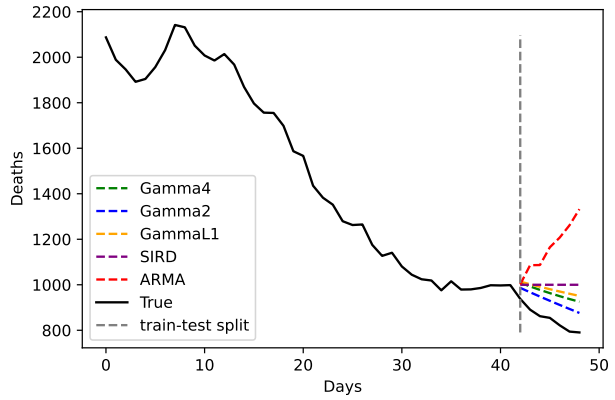
In terms of weaknesses of Gamma Model, we first mention that our model relies on very rigid assumptions regarding the delay of information. Specifically, we assume that infections within the V segment are observed after exactly $\Delta_1 = 11$ days, and that deaths happen exactly $\Delta_2 = 18$ days after the infection. These assumptions likely cause a considerable proportion of the error in prediction, as the time it takes to get hospitalized or pass away from the disease has a good deal of variance.

Moreover, our model assumes there is a perfect split (done a-priori) into two population segments V, U, in terms of whether a hospitalization would occur or not. However, such a split is not realistic, as hospitals' policies (depending on symptoms) regarding of what patients they admit can differ a lot in the course of the pandemic, because during peaks of infection, resources such as hospital beds are scarce. Additionally, as new treatments against the disease have been developed, fewer people might need hospitalization to fight the disease, so the demarcation between U-types and V-types is not stable throughout the entire period of the disease. This issue also comes up in our assumption that the probability of death given infection is constant. Novel treatments, as well as more patient time during periods of low case count, can help reduce the mortality rate.

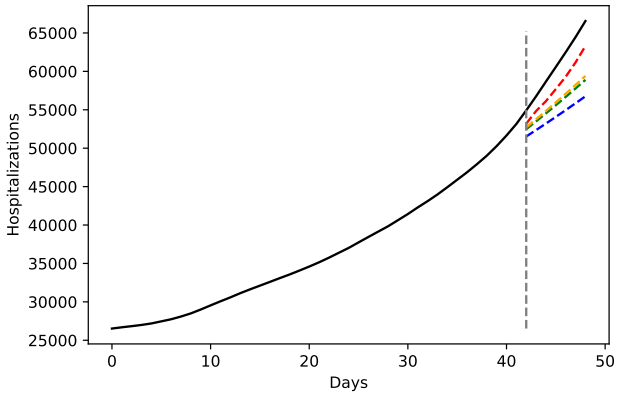
Finally, another drawback of Gamma Model is that, due to the difference in Δ_2, Δ_1, even though we receive the ground truth values of $V_{t'}$ for $t' \in \{t - \Delta_2, ..., t - \Delta_1\}$, we cannot, at time t, compute $U_{t'}$ for $t' \in \{t - \Delta_2, ..., t - \Delta_1\}$. Hence, after estimating $\gamma_{t-\Delta_2}$, we use this value to propagate forward the values of $\hat{V}_{t-\Delta_2+1}, \hat{V}_{t-\Delta_2+2}, ...$, even though the values $V_{t-\Delta_2+1}, V_{t-\Delta_2+2}, ..., V_{t-\Delta_1}$ are known to us. Not being able to use the ground truth values of hospitalizations for the last $\Delta_2 - \Delta_1$ days inevitably affects the quality of our predictions, as we essentially propagate predictions forward 14 days instead of just 7. This is one of the main reasons why, as it can be seen from Figure 1, some of our hospitalization predictions diverge from the test set.

Table 2: Total time to fit all $B = 25$ models for each model method with United States time-series.

Model	Fit Time (s)
ARMA	67.90
Gamma4	5.20
GammaL1	7.48
Gamma2	3.19
SIRD	143.50



(a) Prediction of deaths over time.



(b) Prediction of hospitalizations over time.

Figure 1: Example of the prediction in the United States. Models were trained on the first 42 days and then made predictions for the following 7 days.

5 Conclusion

We have developed a new model, *Gamma model*, that captures the true infection rate of the COVID-19 pandemic, in order to predict hospitalizations and deaths in the near future. Our model does not rely on the number of confirmed cases which can be inaccurate or deficient. Instead, we split the population into two groups, according to the risk of getting severe symptoms, and estimate their size based on deaths and hospitalizations. We compared our model with the L1-regularized autoregressive moving average (ARMA) and the SIRD model, which have been widely used in predicting the trajectory of the COVID-19 pandemic. We have shown that the Gamma model can match the performance of ARMA and SIRD at a much lower computational cost.

There are several avenues that we could explore in future work. First, we consider adapting the model so as to incorporate the 7-day unused ground truth hospitalization data (instead of predicting them). Additionally, we can extend our model with vaccination information.

References

Alzahrani, S. I., I. A. Aljamaan, and E. A. Al-Fakih (2020). "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions". In: *Journal of Infection and Public Health* 13.7, pp. 914–919. ISSN: 1876-0341.

Ceylan, Z. (2020). "Estimation of COVID-19 prevalence in Italy, Spain, and France". In: *Science of The Total Environment* 729, p. 138817. ISSN: 0048-9697.

Chowdhury, R. et al. (May 2020). "Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries". English. In: *European Journal of Epidemiology* 35.5.

Cordelli, E., M. Tortora, R. Sicilia, and P. Soda (2020). "Time-Window SIQR Analysis of COVID-19 Outbreak and Containment Measures in Italy". In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 277–282.

Gribaudo, M., M. Iacono, and D. Manini (2021). "COVID-19 Spatial Diffusion: A Markovian Agent-Based Model". In: *Mathematics* 9.5. ISSN: 2227-7390.

Hasell, J., E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie (2020). "A cross-country database of COVID-19 testing". In: *Scientific data* 7.1, pp. 1–7.

Hoseinpour Dehkordi, A., M. Alizadeh, P. Derakhshan, P. Babazadeh, and A. Jahandideh (2020). "Understanding epidemic data and statistics: A case study of COVID-19". In: *Journal of Medical Virology* 92.7, pp. 868–882.

Kermack, W. O., A. G. McKendrick, and G. T. Walker (1927). "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772, pp. 700–721.

Mandal, M., S. Jana, S. K. Nandi, A. Khatua, S. Adak, and T. Kar (2020). "A model based study on the dynamics of COVID-19: Prediction and control". In: *Chaos, Solitons & Fractals* 136, p. 109889. ISSN: 0960-0779.

Menda, K., J. De Becdelievre, J. Gupta, I. Kroo, M. Kochenderfer, and Z. Manchester (2020). "Scalable Identification of Partially Observed Systems with Certainty-Equivalent EM". In: *ICML*. PMLR, pp. 6830–6840.

Nesterov, Y. (2020). "Online analysis of epidemics with variable infection rate". In: *arXiv preprint arXiv:2007.11429*.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pourghasemi, H. R., S. Pouyan, Z. Farajzadeh, N. Sadhasivam, B. Heidari, S. Babaei, and J. P. Tiefenbacher (2020). "Assessment of the outbreak risk, mapping and infestation

behavior of COVID-19: Application of the autoregressive and moving average (ARMA) and polynomial models". In: *medRxiv*.

Sen, D. and D. Sen (2021). "Use of a Modified SIRD Model to Analyze COVID-19 Data". In: *Industrial & Engineering Chemistry Research* 60.11, pp. 4251–4260. DOI: 10.1021/acs.iecr.0c04754.

Weinberger, D. M. et al. (Oct. 2020). "Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020". In: *JAMA Internal Medicine* 180.10, pp. 1336–1344. ISSN: 2168-6106.

Yousaf, M., S. Zahir, M. Riaz, S. M. Hussain, and K. Shah (2020). "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan". In: *Chaos, Solitons & Fractals* 138, p. 109926. ISSN: 0960-0779.

Zou, D., L. Wang, P. Xu, J. Chen, W. Zhang, and Q. Gu (2020). "Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States". In: *medRxiv*. DOI: 10.1101/2020.05.24.20111989.