Models and Optimization for Forecasting COVID-19

Andrei Graur, Arec Jamgochian, Bernard Lange, Kyriakos Lotidis

Introduction

COVID-19 pandemic has led to more than 3 million deaths and 150 million cases worldwide. The exponential rate of new infections has exposed the weaknesses of pandemic preparedness. We aim to obtain an accurate estimate of the infection rate.

Methodology

1 Gamma model

Gamma model captures the dynamics of the COVID-19 spread. We define:

- V - vulnerable population: will end up in a hospital case in Δ_1 days when infected
- U - non-vulnerable population: will not require hospitalization
- 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 individuals that are actively infected during day t
  - U_t - number of type U individuals that are actively infected during day t

Assumptions:

1. $h_t = V_{t-\Delta_1}$. Every case among type V individuals is detected in Δ_1 days.
2. $d_t = p(T_{t-\Delta_2} - T_{t-\Delta_2-1})$. Death occurs in exactly Δ_2 days.
3. For each day t, a matrix $\mathbf{\Gamma}_t \in \mathbb{R}^{2\times 2}$ exists such that

$$\begin{bmatrix} V_{t+1} \\ U_{t+1} \end{bmatrix} = \mathbf{\Gamma}_t \begin{bmatrix} V_t \\ U_t \end{bmatrix} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2 \begin{bmatrix} V_t & 0 \\ 0 & U_t \end{bmatrix})$$

Objective: The task is to find the solution $\mathbf{\Gamma}$ to the following convex problem:

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

where L1 term penalizes daily change in $\mathbf{\Gamma}_t$ to enforce smoothness, with weight λ.

.

Solver: We employ the subgradient descent. We set the number of iterations to be 10^5, and stepsize to $\alpha_k = \frac{\alpha}{\sqrt{k}}$ for $\alpha = 0.1$. In our experiments, CVXPY was not converging in the finite number of iterations.

Prediction: 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 $\mathbf{\Gamma}_{t-\Delta_2}$ as follows:

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

An important aspect to note is that, we must predict $\hat{V}_{t-\Delta_2+1:t-\Delta_1}$ as described above as well, and then proceed to predict $\hat{V}_{t-\Delta_1+1:t-\Delta_1+K}$. so for $K = 7$ days ahead of time t. We predict the number of deaths to be $p \cdot (\hat{V}_{d-\Delta_2} + \hat{U}_{d-\Delta_2})/14$.

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

Fig. 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.

2 Regularized Autoregressive Moving Average (ARMA)

We use L steps of hospitalizations and deaths to regress to next step hospitalizations and deaths.

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

We fit $\mathbf{\Theta}$ by minimizing the objective

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

This optimization objective can be solved efficiently with coordinate descent.

3 SIRD model

A Susceptible-Infectious-Recovered-Deceased (SIRD) model of the infectious disease. The population is divided into: susceptible $\boldsymbol{S(t)}$, infectious $\boldsymbol{I(t)}$, recovered $\boldsymbol{R(t)}$, deceased $\boldsymbol{D(t)}$, and confirmed cases $\boldsymbol{C(t)}$. The model is defined as follows:

$$\begin{gathered} \frac{dS}{dt} = -\frac{\beta I S}{N}, \qquad \frac{dI}{dt} = \frac{\beta I S}{N} - \gamma I - \mu I, \qquad \frac{dR}{dt} = \gamma I \\ \frac{dD}{dt} = \mu I, \qquad \frac{dC}{dt} = \frac{\beta I S}{N} \end{gathered} \tag{3}$$

where β, γ, and μ are constants that we must optimized for, and only C and D are observed. 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

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

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), $\boldsymbol{\theta}$ contains model parameters (β, γ, and μ), and w_t and v_t are process and observation noises. We use Certainty-Equivalent Expectation Maximization (CE-EM) (Menda et al., 2020). CE-EM iteratively performs a two-step procedure—the E-step holds $\boldsymbol{\theta}$ constants and infers the unobserved state variable $\mathbf{x}$, while the M-step optimizes for $\boldsymbol{\theta}$. It only maintains the most likely estimate. It uses Nelder-Mead method to fit θ and Gauss-Newton to estimate state variables.

Experiments

4 Dataset

We use the *Our World in Data* Covid-19 dataset (Hasell et al., 2020). It consists of statistics reported daily for a number of countries. We extract time-series of new cases, total hospitalizations, and new deaths reported daily in the United States, the United Kingdom, and Italy. We randomly select $B = 25$ sets of 42-day input series to use for training, and the $K = 7$ following days to use for testing. We report the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE).

5 Results

We report the metrics on unseen statistics in Table 1. In Figure 2, we visualize a single 7-day death and hospitalizations predictions for all models.

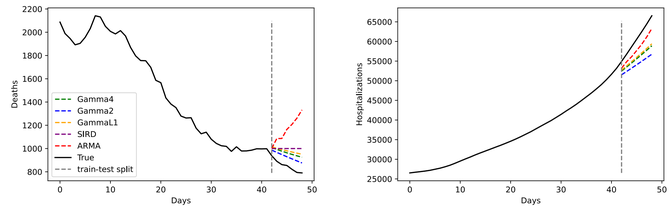

Fig. 2: Prediction of deaths (left) and hospitalizations (right) over time.

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

Fig. 3: Total time to fit all $B = 25$ models for each model method with United States time-series.

Discussion and Conclusion

No model uniformly outperforms the others. Furthermore, we notice by examining the MAPE that at best, we can only expect our models to achieve $8-15\%$ accuracy in predicting deaths over 7 days. We observe that there is a large gap in model fitting times between the three models. Our Gamma model can match the performance of ARMA and SIRD at a much lower computational cost.

References

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.

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.