Supplementary Material for ”Survival-Convolution Models for Predicting COVID-19 Cases and Assessing Effects of Mitigation Strategies”

Qinxia Wang*, Shanghong Xie*, Yuanjia Wang^{*,1}, Donglin Zeng^{†,1}
*: Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA ;
†: Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

April 15, 2020

Model Estimation and Inference

We propose a parsimonious survival-convolution model for predicting key statistics of COVID-19 epidemics (daily new cases). For the survival function of presenting COVID-19 symptoms, previous work[1] indicates that the incubation period for COVID-19 has an average of 5·2 days and that the longest time was reported up to 21 days. Thus, we assume that the survival function for COVID-19 symptom onset, $S(k)$, follows an exponential distribution with mean 5·2 truncated at 21 days. We model the infection rate $a(t)$ as a non-negative piece-wise linear function (linear spline and assume $a(t) \geq 0$). For China and South Korea, $a(t)$ is given as follows:

$$a(t) = \begin{cases} a_0^+ & t < t_1 \\ (a_0 + a_1(t - t_1))^+ & t \geq t_1 \end{cases}, \tag{s1}$$

where $x^+ = \max(x, 0)$ and t_1 is the calendar time of reporting the first case. That is, before the first case is reported, the public is unaware and the infection is latent, so the infection rate is assumed to be a constant; however, once the first case is reported, the public is alerted and various response strategies are gradually introduced and take effect, so that we expect the infection rate will decrease (i.e., $a_1 \leq 0$). In this simple model, there are three parameters that will be estimated from data, including t_0 (the date of the first case), a_0, and a_1.

When a massive public health intervention (e.g., nation-wide lockdown) is introduced at some particular date, we further add an additional linear function after this date and introduce a new slope parameter. Thus, the change in the slope parameters before and after an intervention reflects its effect on reducing the rate of decline in disease transmission (i.e., “flattening the curve”). Furthermore, since the intervention effect may diminish over time, we introduce another slope parameter two weeks after intervention to capture the longer-term effect. Thus, for Italy and US we place additional knots at t_2 (the date of national lockdown for Italy and the declaration of national emergency for US) and another knot at t_3 (two weeks after t_2). Therefore $a(t)$ is modeled as:

$$a(t) = \begin{cases} a_0^+ & t < t_1, \\ (a_0 + a_1(t - t_1))^+ & t_1 \leq t < t_2, \\ (a_0 + a_1(t_2 - t_1) + a_2(t - t_2))^+ & t_2 \leq t < t_3, \\ (a_0 + a_1(t_2 - t_1) + a_2(t_3 - t_2) + a_3(t - t_3))^+ & t \geq t_3. \end{cases} \tag{s2}$$

We let θ denote all parameters and let $Y(t;\theta)$ denote the predicted new case number at day t for a given θ using recursive equations in (1) and (2) in the main manuscript. To estimate θ, we divide the reported daily new cases

[1] Correspondence to: Dr. Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA (yw2016@cumc.columbia.edu) and Dr. Donglin Zeng, Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA (dzeng@email.unc.edu)

into training data to fit the model and testing data for validation. Denoted by $Y_o(t_1), Y_o(t_1+1), Y_o(t_1+2), ..., Y_o(t_2)$, the training data consisting of the daily new cases reported from the date of the first reported case, t_1, to the last date in the training set, t_2. We minimize the following loss

$$\sum_{t_1 \leq t \leq t_2} \left[\sqrt{Y_o(t)} - \sqrt{Y(t;\theta)} \right]^2 \tag{s3}$$

to estimate θ. The square-root transformation is applied to the daily cases since this transformation is known to be a variance stabilizing transformation for Poisson counts. Computationally, we perform grid search of t_0 and for each t_0, we apply a gradient-based optimizer with adaptive learning rate (i.e., *Adam*[2]) to obtain other parameters. The algorithm is implemented in Tensorflow[3]. We let $\hat{\theta}$ be the minimizer of (s3). With $\hat{\theta}$, we can use equations (1) and (2) in the main manuscript to predict any new daily cases in future dates. Furthermore, by comparing the estimated $a(t)$ (and correspondingly, R_t) before and after an intervention occurs, we can estimate the intervention effect in terms of the change of infection rates under the regression continuity design.

For statistical inference such as obtaining confidence intervals of predicted numbers or estimated intervention effects, we assume that the standardized residuals, $[Y_o(t) - Y(t;\theta)] / \sqrt{Y(t;\theta)}$, are exchangeable. Thus, permutation method can be used. We permute the estimated residuals and reconstruct observed cases by adding permuted residuals multiplied by the square-root of the observed case numbers. We repeat this process 500 times and re-analyze each set of permuted data to yield a set of estimates for θ, the corresponding set of predictions for $Y(t;\theta)$ and estimated intervention effects. We obtain 95% confidence intervals using quantiles of the set of estimates.

References

1 Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine 2020;382:1199–1207.

2 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint 2014;arXiv:1412.6980.

3 Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015. Software available from https://www.tensorflow.org.