Data Driven Predictive Modeling of Infectious Disease Spread using the SIR Model

Somenath Biswas, and A. V. Sreejith

IIT Goa

May 2020

1 Introduction

Currently the entire world is going through a nightmare and everyone wishes to know how long this nightmare will last, and what its toll will be. A number of academic groups across the world are carrying out data driven predictions in answer to these questions; *Predictive Monitoring of COVID-19* from the SUTD Data Driven Innovation Lab of the Singapore University of Technology and Design is one such example. (`https://ddi.sutd.edu.sg/`, links to other such efforts can be found there.) Most of these efforts make use of a mathematical model of infection spread, usually the SIR model, or one of its variants like SEIR or SIRD etc., and then learn certain model parameters from the publicly available data of the infection in a region to make predictions for the region. Learning model parameters is done by, what is called in statistics and machine learning parlance, *regression*. The most used regression method is *linear regression* which essentially is taking the projection of a vector on to a linear space. The purpose of this note is to explain the SIR model and how the evolving infection data is used to fix the SIR model parameters. We refer to [He00] for a survey of the SIR model, its variants, and how effective such a model is in modeling certain past epidemics. Any standard text on linear algebra, e.g., [St06] explains what is the projection of a vector on to a linear space.

2 The SIR model

Our description of the model follows *The SIR Model for Spread of Disease– The Differential Equation Model* by David Smith and Lang Moore [SM04]. SIR is the acronym for Susceptible-Infected-Recovered. The model aims at capturing quantitatively the spread of a new infectious disease in a closed community of N individuals; the disease being new implies

that no one in the community has a priori immunity from the disease, infectious implies that an infected individual can potentially infect any uninfected one coming in contact with the infected individual, and the community being closed means that there is no movement in and out of the community, and therefore, throughout the life of the disease, the population remains the same, namely, N. Unfortunately, some individuals may succumb to the disease, and yet the assumption that the population remaining the same does make sense in terms of the model. We will point out the reason later.

The model uses time, t, as the independent variable, usually measured in the unit of days, and has three dependent variables:

- $S(t)$: The number of of individuals *susceptible* to the disease at time t,
- $I(t)$: The number of individuals already *infected*, and
- $R(t)$: The number of individuals who have *recovered* from the disease.

The model assumes that once recovered, an individual develops the required immunity which protects him from getting infected again. Also, having recovered, a person will not infect anyone else. The model includes the individuals who succumb to the disease amongst the recovered set, as a deceased individual, like a recovered person, will not catch the disease again, nor will he infect anyone else. As the model aims to study the spreading process, this (seemingly bizarre) decision of including the deceased amongst the recovered set is justifiable. [1]

We therefore have that

$$S(t) + I(t) + R(t) = N \tag{1}$$

for all time t.

The above implies

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0 \tag{2}$$

The model defines what these three derivatives are making use of two parameters b and k, and these parameters are defined as follows:

- b: This parameter is a constant which is defined to be the fixed number of contacts that an infected individual has per day. Defining $s(t)$ to be $S(t)/N$, namely, the proportion of the population which is susceptible, of the b contacts per day made by an infected individual, $bs(t)$ will be the number of individuals who are susceptible in expectation, making the convenient *uniformity* assumption that the contacts of the infected person is similar in characteristics to that of the entire population. A slight rewrite: This parameter is a constant which is defined to be the fixed number of contacts that an

[1] We note that the model can be easily modified to have the deceased as a separate category as well.

infected individual has per day. Defining $s(t)$ to be $S(t)/N$, namely the proportion of the population which is susceptible, $bs(t)$ will be the number of susceptible individuals in contact with an infected person. We make the convenient *uniformity* assumption that the contacts of the infected person are similar in characteristics to that of the entire population.

- k: It is assumed that a fixed fraction of the infected individuals joins the recovered group each day. k denotes this fixed fraction.

Using the definitions of $S(t), R(t)$, and $I(t)$, and $s(t)$, and the definitions of the two parameters, we get the following:
Rate of change of the number of susceptible:

$$\frac{dS(t)}{dt} = -bs(t)I(t) \tag{3}$$

Rate of change of the number of recovered:

$$\frac{dR}{dt} = kI(t) \tag{4}$$

Using Equations 2, 3, and 4, we get the rate of change of the number of the infected:

$$\frac{dI}{dt} = bs(t)I(t) - kI(t) \tag{5}$$

As $s(t)$ monotonically decreases with time, whereas k remains a constant fraction, the rate of change in the number of infected will decrease with time, and the graph of $I(t)$ will have a bell-like appearance. Initially, for a certain interval, $I(t)$ will keep on increasing till it peaks, and then $I(t)$ will decrease with time, and eventually the infection will die out.

3 Learning best values for model parameters from data

We may think of the above (5) as a continuous dynamical system. Making first-order Euler discretization of the above continuous system, we get a discrete dynamical system $(I_n)_{n\geq 0}$. For $n = 1, 2, \ldots$

$$I_n = I_{n-1} + (bs_{n-1}I_{n-1} - kI_{n-1})\Delta t \tag{6}$$

where I_n denotes the number of infected at the nth time step, usually the time step is a day. Setting Δt to 1 as we study the evolution of the dynamical system in terms days, we get:

$$(s_{n-1}I_{n-1})b - I_{n-1}k = I_n - I_{n-1} \tag{7}$$

for all $n \geq 1$. On day 0, only a tiny fraction ϵ of the population is infected, that is, $I_0 = \epsilon N$, and $s_0 = 1 - \epsilon$.

Suppose, we observe the data of the number of the infected people, and the number of the recovered people for every day for the first $m+1$ days, starting from the 0th day. We would like to use this data for estimating b and k. Using the relation (7), the data[2] give rise to m equations in two unknowns b and k:

$$c_{j_1} b + c_{j_2} k = w_j, 1 \leq j \leq m \tag{8}$$

where cj_1 is $s_{j-1}I_{j-1}$, c_{j_2} is $-I_{j-1}$, and w_j is $I_j - I_{j-1}$, that is, the new cases found on the jth day.

Let y denote the column vector of the unknowns $\begin{bmatrix} b \\ k \end{bmatrix}$
Then, in the matrix form, the above m equations give us a system of equations

$$Cy = w$$

The system above will most surely be inconsistent, we cannot simply solve for y. Instead, one finds a $\hat{y}$ in the column space of C such that with this $\hat{y}$, $\|C\hat{y} - w\|$ is minimized. The rationale is that, so far as the observed data is concerned, the vector $C\hat{y} - w$ is the error for the choice of $\hat{y}$ as the 'solution' of the system of equations, and therefore, the norm of the error vector is minimized for the best result. Geometrically, it is evident that $\|C\hat{y} - w\|$ will be minimized when $C\hat{y}$ is the projection of w on to the column space of C, that is, when the vector $C\hat{y} - w$ is orthogonal to the column space of C, which will happen when $C\hat{y} - w$ is orthogonal to each column of C. In other words, we seek a $\hat{y}$ for which

$$C^T(C\hat{y} - w) = 0$$

That is,

$$C^T C\hat{y} = C^T w \tag{9}$$

Claim 1 *If columns of C are independent, then C^TC is invertible.*

A geometric way to see why C and C^TC has the same null space. Let $C^T(Cx) = 0$ for some x and let $z = Cx$. Clearly z is in the column space of C. On the other hand $C^Tz = 0$ Therefore, z is perpendicular to all the rows of C^T. In other words, z is perpendicular to the column space of C. Combining we get, z is both perpendicular to the column space and is in the column space. This can only happen if $z = 0$ which implies $Cx = 0$. The claim follows from the observation that both C and C^TC have the same null space, i.e., for all x, $Cx = 0$ iff $C^TCx = 0$. Clearly, if $Cx = 0$, then $C^T(Cx)$ too will be 0. On the other direction, let $C^TCx = 0$. Multiplying both sides of the equation by x^T, we get $x^TC^TCx = 0$, This can be re-written as $(Cx)^TCx = 0$, which implies $Cx = 0$. Since columns of C is assumed to be

[2]We need the data for both the infected and the recovered to fix s_p, for $p \geq 1$.

independent, null space of C is the trivial one, namely, $\{0\}$. C^TC too then has only 0 in its null space, therefore, C^TC is invertible. Using the invertibility of C^TC, we get from (9):

$$\hat{y} = (C^TC)^{-1}C^Tw$$

This $\hat{y} = \begin{bmatrix} \hat{b} \\ \hat{k} \end{bmatrix}$ is what we learn from data as the best values for parameters b and k. We note that in practice the two columns of C will surely be found to be linearly independent, as the two entries of the two columns of C of a row make use of the infection data of a particular day, and the infection spread over days is certainly a nonlinear process.

4 An illustration

5 Discussion

How good is the SIR based predictive modeling? Clearly, the SIR model makes certain assumptions which do not hold in practice: the population is never closed entirely– there will certainly be some movements in and out of a region of interest. The uniformity assumptions in the model are also not realistic– usually a child makes more contacts per day than an adult does, and as different parts of the region of interest have different population densities, their contact numbers will also be different. Having a fixed k is also not realistic– different classes of people will have different recovery rates– and also evolving treatment regimes are likely to make k a varying entity with time. Further, not all infected cases are detected, therefore, the data used for fixing the model parameters are also flawed. The model does not take into account societal interventions like lock downs, quarantine, etc. In many cases, as in India, testing becomes more extensive with the progress of time, therefore, giving equal status to the data of each day may itself be faulty. Extensions of the SIR model that address some of these issues have been proposed, we refer to [ABDSS20] for a recent survey.

At the same time, the very simplicity of the basic SIR model is also its greatest strength– in the face of many uncertainties, the uniformity assumption is quite reasonable as there is likely to be an averaging effect. As [He00] shows, the basic SIR model has been quite successful in modeling many past infections. In the case of the current pandemic, the SIR predictive model predictions are not as catastrophic as the scenarios that some people have come up with. This, if nothing else, gives us at least the courage to face the future with some amount of equanimity!

References

[ABDSS20] Abadie, Alberto, Paolo Bertolotti, Ben Deaner, Arnab Sarkar, and Devavrat Shah, Epidemic Modeling and Estimation, Memo of the IDSS COVID-19 Collabora-

tion Project, Institute of Data Science and Society, MIT, 2020.

[He00] Hethcote, Herbert W., The Mathematics of Infectious Diseases, SIAM Review, Vol. 42, No. 4 (Dec., 2000), pp. 599–653, 2000.

[SM04] Smith, David and Lang Moore, The SIR Model for Spread of Disease, Convergence, Mathematical Association of America, December 2004.

[St06] Strang, Gilbert, *Linear Algebra and its Applications*, 4th Edition, Thomson Brooks/Cole, 2006.