Description of the Model and the Parameters in "Inference on the dynamics of COVID-19 from observational data"

1 Model and assumptions

We assume that the dynamics of the COVID pandemic can be described in terms of the evolution of a few observable and unobservable compartments. The observable compartments in our model are:

- Q_t = number of asymptotic individuals who have are currently in quarantine (i.e., they have been tested positive but do not show any significant symptom, and so are not hospitalized)
- H_t = number of people who are in hospital at time t.
- D_t = number of deaths due to the disease up to time t.
- C_t = number of confirmed cases up to time t.
- T_t = number of tests up to time t.

The unobservable compartments are:

- A_t = number of infected but asymptomatic individuals at time t.
- S_t = number of susceptible individuals in the population at time t.

We also define compartments R_t^Q and R_t^H, which represent the number of recoveries from quarantine and hospitals, respectively, up to time t. Analogously, we define R_t^A to be the number of individuals who recover from the asymptomatic compartment without either being tested positive. Clearly, R_t^Q and R_t^H are observables (though we may not always have reliable data on these), while R_t^A is unobservable. This leads to the following definitions for the recovered compartment.

- $R_t = R_t^A + R_t^H + R_t^Q$ = the total number of recovered individuals up to time t,

which is an unobservable quantity is reality. The reported number of poeple recovered from the disease is only a subset of R_t, and is given by

- $R_t^{reported} = R_t^H + R_t^Q$,

In addition, our model depends on κ_t = the current state of social distancing, which is expressed as a fraction, with the value 1 being normal activity, and 0 being indicative of complete lockdown, and hence no interaction among individuals. In many cases, κ_t is not fully observable. However, one might use a surrogate value of κ_t based on data collected by internet service providers such as Google from the usage of smartphones.

For simplicity, we make the following assumptions:

A1 Only an asymptomatic individual who is not either in quarantine or in hospital can transmit the disease to a susceptible individual.

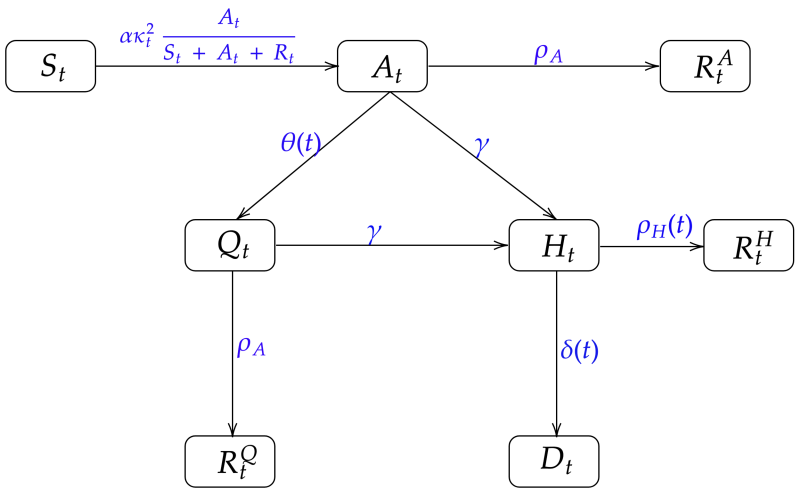

Figure 1: Flowchat showing the different compartments and rates in the propagation of the pandemic, as considered in the model. S_t, A_t, H_t, Q_t, D_t are the number of susceptible, infected, hospitalized, quarantined, and deceased people at time t respectively. R_t^Q, R_t^H, R_t^A represent the recovered population from quarantined, hospitalization, and infected but asymptotic stages respectively. The rate parameters are as described in Section 1.1.

A2 People who recover from the disease are immune from subsequent infection.

A3 False positive rate for the test is negligible, so that if somebody is confirmed to be positive, then he/she is assumed to be infected.

A4 Anybody who shows significant symptoms, whether being in quarantine or not, is immediately hospitalized, and is tested to be positive.

A5 There is no effective treatment regime for the asymptomatic individuals, and so

they recover or turn symptomatic at the same rate regardless of whether they are tested positive (and hence quarantined) or not.

We can adapt a modification of the model if **A2** is violated by adding a fraction of people from the recovered compartment to the susceptible compartment. These assumptions are, in general interpretable and quite realistic, however, they inherently depend on the people behaving responsibly throughout the pandemic. An efficient market research could be employed to validate the assumptions regarding peoples behavior during the pandemic.

1.1 Statistical model

We assume a Poisson process model for describing the dynamics. According to this model, conditionally on the current values of different compartments (collectively denoted by $\mathcal{F}_t$), the quantities ΔA_t, ΔR_t, ΔD_t, ΔH_t and ΔC_t follow Poisson distributions that depend on $\mathcal{F}_t$ only through a time-varying rate parameter. Accordingly, the proposed model can

be expressed through the following set of equations:

$$\begin{aligned}
\Delta S_t &= -\Delta A_t && (1)\\
\mathbb{E}(\Delta A_t|\mathcal{F}_t) &= -(\theta(t)+\gamma+\rho_A)A_t + \left(\frac{S_t}{S_t+A_t+R_t}\right)\alpha\kappa_t^2 A_t && (2)\\
\mathbb{E}(\Delta Q_t|\mathcal{F}_t) &= \theta(t)A_t - (\gamma+\rho_A)Q_t && (3)\\
\mathbb{E}(\Delta H_t|\mathcal{F}_t) &= \gamma(A_t+Q_t) - (\rho_H(t)+\delta(t))H_t && (4)\\
\mathbb{E}(\Delta D_t|\mathcal{F}_t) &= \delta(t)H_t && (5)\\
\mathbb{E}(\Delta C_t|\mathcal{F}_t) &= (\theta(t)+\gamma)A_t && (6)\\
\mathbb{E}(\Delta R_t^A|\mathcal{F}_t) &= \rho_A A_t && (7)\\
\mathbb{E}(\Delta R_t^Q|\mathcal{F}_t) &= \rho_Q(t)Q_t && (8)\\
\mathbb{E}(\Delta R_t^H|\mathcal{F}_t) &= \rho_H(t)H_t && (9)\\
\mathbb{E}(\Delta R_t|\mathcal{F}_t) &= \Delta R_t^H + \Delta R_t^Q + \Delta R_t^A. && (10)
\end{aligned}$$

The parameter ρ_A quantifies the rate of recovery from the asymptomatic compartment. α is the baseline infection rate, in the absence of any social distancing. This means, α is the average number of susceptible individuals who may be infected on any given day by an asymptomatic but infected individual. γ is the rate at which an asymptomatic individual may become symptomatic, and is considered to be the same whether the individual is free or in quarantine. $\rho_H(t)$ is the rate at which people recover from the hospitalized compartment. This parameter, as well as $\delta(t)$ = rate of death from the hospitalized compartment, are both time-varying to reflect the changing levels of effectiveness of treatment regimes over time. Observed that, under the assumption **A5** we have $\rho_Q(t) \equiv \rho_A$.

The function $\theta(t)$ is not observed and it can be described as the *confirmed fraction (CF)*, since it has the interpretation as the fraction of currently asymptomatic individuals who are detected through testing. Below, we shall relate it to the *testing factor (TF)*

F_t, defined as the number of new tests divided by the current number of hospitalization: $F_t = \Delta T_t / H_t$. Indeed, we express $\theta(t)$ as

$$\theta(t) = \phi(t) F_t = \phi(t) \frac{\Delta T_t}{H_t}, \tag{11}$$

whereby $\phi(t)$ may be referred to as the *testing efficiency (TE)*. Clearly, $\phi(t)$, and hence $\theta(t)$, are nonnegative functions. Also, the parameters α, γ, ρ_A, $\rho_H(t)$ and $\delta(t)$ are nonnegative.

It is also of note, that the rate at which a susceptible individual turns infected (asymptomatic) can be approximately represented as $\alpha\kappa_t^2$. As such the fraction $\frac{A_t}{S_t+A_t+R_t} \approx 1$ throughout the pandemic. This is a prudent assumptions for all practical demonstration of our method, since mitigation measures are implemented in most affected places or countries to contain the spread of the disease, rather than waiting for herd immunity to be achieved. As a consequence, at any given time, the number of asymptomatic but infected people is much lower as compared to the susceptible population.

In reference to (6), two sources of confirmed cases are considered - one from the number of infected (symptomatic or asymptomatic) people who are detected correctly and quarantined, the second from the number of people who turn symptomatic and require hospitalization. ΔC_t is the reported number of newly infected people at time t. We observed that, at any given time, the number of infected but asymptomatic people is unknown. Hence, the number of reported new infections at time t will always be an under-representation of the actual number of new infections. This is due to the fact that the latter considers the total of the newly infected symptomatic and asymptomatic people at time t. Our model is able to discern the unobservable trajectory of true new (also cumulative) infections, which is highly useful in assessing the number of possible spreaders (asymptomatic but infected peopulation) in a particular city, state, or country.

In the real data application of our model, we will discuss further about this discrepancy between the estimated value of the true quantity and the reported number of the infected people. The core of our estimation strategy is to interpret the relationships (2)–(10) as appropriate regression problems, considering the left hand sides of the equations as the responses the right hand sides as predictors, respectively. The corresponding parameters will be estimated following a local linear regression method.

The precise details of the estimators depend on the kinds of data that are available, as well as certain specifics of the model assumption, such as whether a parameter is considered fixed or time-varying. As a general rule, we use local regression (linear or nonlinear) methods for estimating the time-varying parameters, while *profiling* over the time-independent parameters.