Contents

1 Introduction

Since its outbreak in late 2019 in Wuhan, China the COVID-19 virus has spread all over the world and has challenged the social and health systems of many countries. The pandemic has also triggered a huge interest from all disciplines of science and countless researchers have contributed to the better understanding of the characteristics of this novel disease. An astonishing number of more than 29000 COVID-19 related contributions is available on the WHO website and all of these articles where published after the first of January 2020.

Those contributions cover an extremely wide range of aspects of the pandemic: from the behavior of the virus on a microscopic level and it's lifespan on different surfaces to the impact of the lockdown on the psychology of the

population. One very important aspect of the pandemic is obviously the understanding of it's evolutions and consequently it's modeling and forecasting. Modeling is very important because it is the only way for decision-makers to implement the right policies at the right time. A robust model can help to choose what kind of measures to implement, like school closing or travel ban and also when to stop these measures.

The goal of this work is to explore the usage of data-driven system identification tool for COVID-related data. The most common way of modeling a disease is through compartmental models. They divide the population into different compartments like susceptible (not yet infected by the disease), exposed (infected but not yet infectious), infected, recovered. The interaction processes between the different compartments are governed by a set of simple ODEs. This framework was introduced almost a century ago and is flexible enough to model most of the effects and controls of the disease. The classic epidemiological approaches will be further introduced in section 2.

In this work we deliberately choose to not base our modeling on these classic techniques and concentrate on the identification of dynamic processes in the evolution of the disease with very little a priori. The different approaches considered in this work all gravitate around an algorithm proposed in [**?**] called *system identification of nonlinear dynamics*. The details about this methodology will be explained in section 3. In a few words, this algorithm allows to identify the dynamics from observed data through a sparse regression and based on a user-defined library of candidate functions. Throughout this work, we will use this algorithm in combination with datasets collected from different sources in order to better understand both the dynamics of the pandemic and the specificities as well as the limits of this algorithm.

This algorithm can allow us to do all kinds of interesting analyses.

2 Classic modeling methods

It was briefly mentioned in the introduction that the most classic way of modeling an infectious disease is based on dividing the population into compartments. The most basic model of this kind is the SIR model that tracks the number of susceptible, infected and recovered people in the sample. It was originally proposed in [**?**] and has become the most widely used epidemiological model discussion about micro and macro levels. This very simple model is formulated by the following set of equations:

$$\begin{aligned}\dot{S} =& -\beta SI \\ \dot{I} =& \beta SI - \gamma I \\ \dot{R} =& \gamma I\end{aligned}$$

where β and γ are the parameters of the disease. $1/\gamma$ is the time a person remains infectious while β quantifies the infectiousness of the disease. The ratio β/γ is usually denoted R_0 and called the basic reproduction value. It is a very important characteristic of the disease because it represents the average number of people a single person is susceptible to infect. All the measures introduced by the governments in order to limit the impact of the disease were targeted at reducing this R_0 value.

These models can be further elaborated in several ways. The first sophistication can be introduced by adding additional compartments like *exposed, quarantined, hospitalized, deceased* to further mimic the real interactions between these different categories of the population. It is clear that COVID-19 has a lot of delay in it's dynamics [**?**] because it can have an incubation time of up to two weeks.

In order to take this into account it might be necessary to add at least an "exposed" compartment. Thus, when infected, a person is not immediately infectious and must stay a few days in the "exposed" compartment. A few examples of more sophisticated SIR models and adapted specifically to COVID-19 can be seen in [**?**] but as mentioned earlier there is a huge number of such adaptations.

Secondly, knowing that the different age groups of the population are affected in very different ways by the disease [**?**], it might be interesting to subdivide the population into the relevant age groups in order to have a more precise control over the parameters. For example the mortality rate is almost equal to 0 for children and is very high for the people above 80 years of age. Age subdivision is also a widely used practice and can be seen in [**?**] just to name a few.

Thirdly, one very classic way of making the models more complex and precise is by subdividing the region of interest into smaller sub-regions and to have a state for each of these sub-regions. For example a country can be subdivided into regions or counties. Some parameters traducing the migrations between these regions can be introduced. One can imagine that this kind of modeling can be very useful for policy-makers to adopt more precise and

local measures. Few examples of such models can be found in [**?**]. These models are usually called *spatial SIRs*.

Finally, there is no limitation to the complexity of the models that can be built on the basis of the SIR models. They can be combined and adapted to a very wide range of problems. An example of such use might include adding the relevant compartments to model the usage of a tracking mobile app that facilitates the isolation of people who have interacted with another infected person.

2.1 Other modeling techniques

The modeling of infectious diseases is however not limited to the aforementioned compartmental models and lots of different approaches from different disciplines exist. Other approaches for modeling the spread of an infectious disease include:

- **Statistical models** are more inclined towards capturing the random nature of the infectious diseases. Thorough reviews of these approaches can be found in [].
- **Gravity models** can be used as standalone models or in conjunction with the compartmental approach and are called "gravity" models because they rely on considering the effect of distance and the size of donor and recipient communities [].
- **Network-based models** are built on the assumption that the spread of human disease follows its specified contact or spreading paths such as transportation or social contact networks [].
- **Agent-based models** are based on the belief that individuals and their mutual differences are the key to understanding the spread of an infection [].

Additionally, hybrid approaches combining different methods are considered in the literature [] as well as more computationally-oriented methods like cellular automata [] A thorough review about the modeling of infectious diseases can be found in [].

3 Sparse identification of nonlinear dynamics

In this section the central algorithm used in this work will be introduced. This algorithm lies on the crossroads between dynamic modeling and machine learning. On one hand this algorithm is able to fit the observed data much like more classic machine learning algorithms would do. And on the other hand it does so by finding governing ODEs in observed data. In theory this allows to model an arbitrarily complex phenomenon while having the interpretability of a set of simple ODEs. We will see that in practice, when dealing with real-world data and complex problems this technique doesn't hold up to these expectations.

Machine learning is an extremely young branch of science as opposed to dynamic modeling, because mathematicians, physicists and engineers have been modeling physical phenomena for centuries. The shift that happened in the recent years is that (1) computational power has significantly increased and now allows for computations of unprecedented complexity and (2) observation data is getting increasingly available as sensor prices go down. The combination of these two tendencies opens new horizons for modeling increasingly complex processes.

It was quickly mentioned in the introduction that this algorithm has two key features: (1) it relies on a user defined set of candidate functions and the result will only be as good as the candidate functions are with respect to the problem; (2) it uses sparse regression (some weights are gradually zeroed-out) in order to find the most parsimonious formula that still fits the data.

The whole methodology to discover the dynamics can be divided into 3 steps. Let's say that we have a series of observations $(\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_m))$ where $\mathbf{x}$ is the state or the vector state of the system of interest and m is the number of observations. Our goal is to find a function f such that:

$$\dot{\mathbf{x}} = f(\mathbf{x}) \tag{1}$$

1. It is more convenient to work with matrices to do the computations so our first step will be to divide the available data into two matrices:

$$X = \begin{bmatrix} \mathbf{x}(t_1) \\ \mathbf{x}(t_2) \\ \vdots \\ \mathbf{x}(t_m) \end{bmatrix} \text{ and } \dot{X} = \begin{bmatrix} \dot{\mathbf{x}}(t_1) \\ \dot{\mathbf{x}}(t_2) \\ \vdots \\ \dot{\mathbf{x}}(t_m) \end{bmatrix}$$

here the dot notation denotes the time-derivative and it can be either directly observed or computed numerically.

2. The second step of this algorithm is the augmentation of the state with the candidate functions. Let's say that we have a set of p candidate functions $(f_1, f_2, \ldots, f_p)$ that we want to use for this identification. What we need to do is to construct a matrix $\theta(X)$:

$$\theta(X) = \begin{bmatrix} f_1(\mathbf{x}(t_1)) & f_2(\mathbf{x}(t_1)) & \cdots & f_p(\mathbf{x}(t_1)) \\ f_1(\mathbf{x}(t_2)) & f_2(\mathbf{x}(t_2)) & \cdots & f_p(\mathbf{x}(t_2)) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(\mathbf{x}(t_m)) & f_2(\mathbf{x}(t_m)) & \cdots & f_p(\mathbf{x}(t_m)) \end{bmatrix}$$

3. Finally we can run the optimization and find the best set of $\xi_1, \xi_2, \ldots, \xi_p$ that minimize the following equation in the least squares sense:

$$\dot{X} = \theta(X) \times \Xi \text{ where } \Xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{bmatrix}$$

   After the best $\xi_1, \xi_2, \ldots, \xi_p$ were identified, the very small ξ (those that are under a certain threshold called *cutoff value*) can be removed from the equation alongside with the candidate function they correspond to. And then the least squares minimization can be ran again with the new subset of candidate functions and weights. After doing this thresholding a few times the algorithm converges.

This is the very general definition of the algorithm, in our case we used almost exclusively polynomial terms of the state as candidate functions. For example if our state is $\mathbf{x} = (s, i, r)$ like in the compartmental models, then our candidate function be: $\mathbf{x} \mapsto 1, \mathbf{x} \mapsto s, \mathbf{x} \mapsto i, \mathbf{x} \mapsto r, \mathbf{x} \mapsto s^2, \mathbf{x} \mapsto si, \mathbf{x} \mapsto sr, \ldots, \mathbf{x} \mapsto r^n$ where n is the maximum degree of the polynomial terms that we consider. If the state of the system is $\mathbf{x} = x$ like if we just tracked the cumulative number of cases in a single country, then the candidate functions would be: $\mathbf{x} \mapsto 1, \mathbf{x} \mapsto x, \mathbf{x} \mapsto x^2, \mathbf{x} \mapsto x^3, \ldots, \mathbf{x} \mapsto x^n$. A more thorough discussion about polynomial terms is held in section 3.1.2.

Another remark is that this algorithm works equally well with an iterative formulation as it does with the differential one like in equation 1. In the

iterative case we would have a series of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ and the goal is still the same but we want f to express the dynamics in an iterative way:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t) \tag{2}$$

This is slightly more convenient because it prevents us from numerically computing a derivative which obviously introduces some errors.

This work was further extended in a second paper [**?**] that is more centered on PDE identification instead of working with ODEs.

3.1 Hyperparameters

If we consider this algorithm as a machine learning methodology then its hyperparameters would be:

1. the choice of the set of candidate functions;
2. the cutoff value.

Both these parameters deserve a discussion.

3.1.1 Cutoff value

We previously mentioned that the cutoff value is the parameter used for thresholding the small weights in the learning process. In fact, this parameter allows us to choose the sparsity of the resulting model and thus choose the required balance between accuracy and complexity of the model.

In practice, due to the low computational cost of this algorithm[1] TODO: run the timing we are able to compute the results for a very wide range of cutoff values and to choose the one that we prefer. Usually the one satisfying the philosophy of "Ockham's razor", in other words the lowest cutoff value that has satisfying fitting of the data. An example of such a plot can be seen on figure **??**.

further discussion

[1] Less than one second for 100 data points and 10 candidate functions on our i7 machine.

3.1.2 Candidate functions

In a simple problem, we just want our candidate function to include the real dynamics. The authors mostly use polynomial terms as candidate functions because the dynamics they identify in the examples are actually polynomial terms of the variables of the state. In the 7 examples of identification that they provide, the candidate functions always include the real dynamics.

In more general cases and more complex problems, we cannot guarantee the same thing. And instead we use polynomials as approximators and hope that they will be expressive enough to capture the evolution. Obviously, it is very important to make this distinction between making a library large enough to include the real dynamics and using polynomials as universal approximators.

Our choice of polynomial basis functions was motivated by the fact that the authors propose their usage, that they are quite simple to use and understand and finally that the study of basis functions is not completely in the scope of this work. Consequently, the question of whether they are the more relevant candidate functions in our case remains not fully answered. Indeed an elaboration was explored by using rational functions instead of polynomials. The full discussion is available in appendix **??** Our main conclusion from this superficial experiment was that with similar maximum degrees, rational functions provided little to no advantage over the polynomial ones in terms of mean squared deviation while requiring a less-convenient non-linear least squares fitting.

4 Data

Data plays a crucial role in our approach because our aim is to base all the modeling on the observation data. In this section we will describe what kind of data we used for our modeling.

4.1 Observing the pandemic

Because this pandemic has no precedents, the metrics people used to track the state of the disease have evolved during the course of the disease. For example, on the date the redefinition of what a "case of infection" is according to the WHO led to a huge spike in new cases in China []. This example

highlights the fact that the observation data in not completely reliable. Furthermore, due to the big differences in healthcare systems in different countries as well as on a regional scale, it can be unfair to compare the numbers. An example of such difference might be the testing policy. Indeed countries around the word have implemented vastly different testing approaches, from very selective testing of only symptomatic people to massive testing like in Germany and Korea. This difference in the number of tests leads to a big difference in the test positivity rate. The higher the test positivity rate the less we are confident about our real understanding of the scale of the epidemic.

In our study, we mainly used the *cumulative number of cases*. Basically, it is the total sum of people that got infected in a given region, its evolution usually looks like a logistic function. Not to be confused with the "I" in an SIR model which accounts for the number of currently infected people. Tracking the evolution of the pandemic through the cumulative number of people is convenient because (1) it is the minimum information we need to have an idea about the state of the disease in the population and (2) even if the quality of the data is not perfect it is the most widely available information (for example, the number of recovered people is missing or is not reliable for a big number of countries).

The most widely used resource for tracking the number of cases worldwide is the repository of the John Hopkins University. They have collected the numbers from all over the world and provided required APIs for all researchers and visualization dashboards to use. They also provide information about recovered cases but as mentioned earlier, the quality of the data is very doubtful.

4.2 Observing the government measures

In our modeling approach we also need to take into account the measures that were taken by the governments to prevent further spreading of the disease like school closing, travel ban, work from home recommendations and alike. For this we relied on the information gathered by a group of researchers from Oxford [**?**]. They tracked the measures that were implemented in most of the countries worldwide and provided a convenient way to retrieve this n information as indicators and also they compiled all of this information into a stringency index. This index is a linear combination of all the indicators that they observed and it's meant to evaluate how strict are the measures implemented by the governments but it has no information about

their effectiveness or performance.

4.3 Other information

Since the adopted methodology of this work heavily relies on the principles of machine learning and data science, we included additional datasets into our modeling because usually, the more relevant information we have, the better the results. For example, in one of our experiments we chose to make a model that would take into account the trajectories of all available countries. In order to inform the model about the present country, we added more that 40 indicators relevant for the pandemic about the country (health, hygiene and demographic descriptors). This information was available to us through the World Bank API [].

5 Modeling

In this section we will introduce the different experiments that we did with the tools and data that we described earlier.

5.1 Cumulative number of cases and forecasting

The most obvious and easy way of using the system identification algorithm with COVID-19 data is to model the number of cumulative cases in a single country. If we define $\mathbf{x}_t$ as the number of cases in the country of interest at day t. Then, we want to find a function f so that:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t)$$

When using polynomial terms as candidate functions we would look for a purely polynomial formulation of the dynamics because there is only one variable in the state. Which can be mathematically written as:

$$\mathbf{x}_{t+1} = \sum_{k=0}^{p} \xi_k \mathbf{x}_t^k$$

where p is the maximum degree of the polynomial terms.

For example we can apply this method to the evolution of the cumulative number of cases in Germany an this is the formula we will end-up with rerun:

$$\mathbf{x}_{t+1} = 4.07 \cdot 10^{-3} + 1.24 \cdot \mathbf{x}_t - 3.33 \cdot 10^{-2} \cdot \mathbf{x}_t^2 + 1.66 \cdot 10^{-3} \cdot \mathbf{x}_t^3 - 3.21 \cdot 10^{-5} \cdot \mathbf{x}_t^4$$

assuming a maximum degree of 4. The figure **??** represents the trajectory of the real data as well as the trajectory of the identified model. One can see that this model fits fairly well the observation data and that the extrapolation looks fairly plausible. Identified models in all countries are available in appendix **??**.

Given that the identified model is extremely simple and has only one variable we cannot retrieve much useful information from it apart from the extrapolation of the number of cases in the next few days. For that matter, the main application of these country-wise models is forecasting. In order to have a better idea of how these very simple identified models compare to more advanced forecasting techniques we compared them to an ARIMA statistical model.

The experiment was build as follows: we identified the dynamics of the cumulative number of cases in 118 countries with maximum degrees of the polynomials ranging from 2 to 7. We kept the last two weeks of available data out of the training set in order to evaluate and compare the models on a one and two weeks forecast horizon. We trained an ARIMA$(1, 2, 1)$ check model on the same training data in each of the countries and then we compared the results. The trajectories in all countries were rescaled so that the last value of the time-series is 1. This allows the models to more easily learn the parameters and to be able to fairly aggregate the mean squared deviation to do quantitative comparisons.

The forecasts of the ARIMA model are fairly similar in terms of mean squared error to the result of the best performing identified system (the one that has the maximum degree that performs best on a given country). In terms of frequency of best results, the ARIMA model seems to have better results (in 55% of the countries ARIMA has the best forecast for both forecast horizons). In terms of errors both methods are also quite similar but the ARIMA model is still slightly better that the dynamic models. A more thorough description of the results of this experiment is available in appendix **??**.

Limits of the experiment The ARIMA forecasting model requires that the time-series verifies some assumptions like stationarity and that there

are no other predictors verify. The statistical model is also not very easy to parametrize and requires the data to be preprocessed for better results. Given that we haven't done all of these manipulations it would not be fair to quantitatively compare the results. In other worlds: there is a big chance that the parameters of the ARIMA model are sub-optimal and thus we conclude that system identification is not likely to give better forecasting results than the statistical model because in the current parametrization it is already worse.

6 Methodologies

7 Remarks