Data-Driven COVID Modeling

Prof. Enrique Zuazua in collaboration with Cyprien Neverov (Intern from IMT Mines Ales)
Chair of Applied Analysis, Alexander von Humboldt-Professorship, Friedrich-Alexander University of Erlangen-Nüremberg

Tuesday 26th May, 2020

Introduction

Exploring and modeling the dynamics of COVID-related data:

- Through sparse system identification of nonlinear dynamics.
- Considering various sets of variables from different datasets.
- Considering different scales of the problem: from country-wise systems to systems including the evolution of the disease in several countries.
- Keeping the modeling purely data-driven, without including any a priori from epidemiological models.

Notation

In this presentation **x** will designate the state of the systems we consider and will always be the number of cumulative cases in a given country.

System identification of nonlinear dynamics

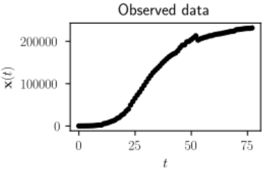
A system identification method proposed by Brunton et al. [1] based on two key features:

- Candidate functions: we have to provide a library of functions susceptible to be relevant for the underlying dynamics.
- Sparse regression: the method is designed to converge to a parsimonious formulation.

This method allows us to find a function f that formulates the ODE that governs the state $\mathbf{x}$ of the system of interest:

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

System identification of nonlinear dynamics



data in

observed state

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

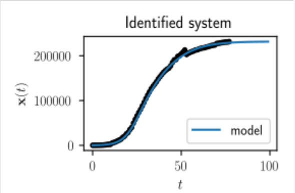
observed or computed time-derivative of the state

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

time derivatives | state augmented with candidate functions | weights

$$\underset{\dot{X}}{\begin{bmatrix} \dot{\mathbf{x}}(t_1) \\ \dot{\mathbf{x}}(t_2) \\ \vdots \\ \dot{\mathbf{x}}(t_m) \end{bmatrix}} = \underset{\theta(X)}{\begin{bmatrix} 1 & \mathbf{x}(t_1) & \mathbf{x}(t_1)^2 & \cdots & \mathbf{x}(t_1)^n \\ 1 & \mathbf{x}(t_2) & \mathbf{x}(t_2)^2 & \cdots & \mathbf{x}(t_2)^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}(t_m) & \mathbf{x}(t_m)^2 & \cdots & \mathbf{x}(t_m)^n \end{bmatrix}} \times \underset{\xi}{\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}}$$

model out



identified ODE: $\dot{\mathbf{x}} = 4.07 \cdot 10^{-3} + 0.24 \cdot \mathbf{x} - 3.33 \cdot 10^{-2} \cdot \mathbf{x}^2 + 1.66 \cdot 10^{-3} \cdot \mathbf{x}^3 - 3.21 \cdot 10^{-5} \cdot \mathbf{x}^4$

The weight matrix ξ is found through a linear least squares optimization. The sparsity is achieved by running the optimization several times and gradually zeroing out the terms that are under a certain cutoff value.

Envisioned directions

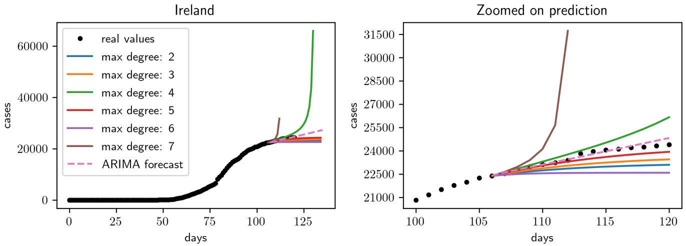
We envisioned several ways of using this system identification tool for COVID-related data:

1. Comparing its forecasting capabilities to classical forecasting techniques in a single-country setting where we only model the cumulative number of cases.
2. Adding external country characteristics as well as information about government measures to model the evolution of cumulative cases in several countries.
3. Studying the relationship between control events like school closures or travel restrictions and the evolution of the number of cases.

Forecasting

We identify the dynamics of the number of cumulative cases **x** in a given country with a maximum degree of polynomial terms ranging from 2 to 7:

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



Result

Regardless of the maximum degree, in more than 55% of the countries, identified models were outperformed by the statistical ARIMA model.

Model for multiple countries

The idea behind this model is to have a single formula that governs the evolution of cases in several countries. For this we provide the model with additional information:

- Indicators about the countries (HDI, total population, and more than 35 other similar health, hygiene and demographics indicators from [2]), a vector $\mathbf{i}_c$ for each country c.
- Information about the government measures through the "Stringency" index from [3] that we can call $h_c(t)$.

We want to identify a function f so that for any country c, at any day t we have:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, h_c(t), \mathbf{i}_c)$$

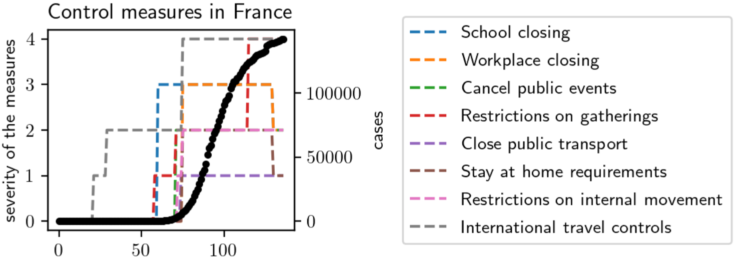
Results

- Identified models overfit to the data: they have good fitting on the countries that it was optimized on and the trajectories do not make any sense on other countries. The models don’t generalize.
- An ODE with polynomial terms might not be a very adapted model for this particular ambitious task.

Impact of the government control

Can we assess the impact of the different control measures ? We try to do this by adding variables about government measures $(h_1(t), h_2(t), h_3(t), \ldots, h_k(t))$:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, h_1(t), h_2(t), h_3(t), \ldots, h_k(t))$$



Result

This approach has not yet been completely explored but we can already observe similar overfitting like in the multicountry model.

Conclusion

- Using system identification with real data is challenging.
- Parameters like maximum degree and cutoff value have a very big effect on the results of the identification and this effect is not very well understood.
- When presented with a big number of variables, identified models tend to overfit.
- Small modifications to the data like rescaling can have a massive impact on the results of the identification.
- The fact that the pandemic is not over yet and the evolutions are still incomplete certainly has some impact on the models we identify.

References I

[1] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016, ISSN: 0027-8424. DOI: 10.1073/pnas.1517384113. eprint: https://www.pnas.org/content/113/15/3932.full.pdf. [Online]. Available: https://www.pnas.org/content/113/15/3932.

[2] *Understanding the Coronavirus (COVID-19) pandemic through data. World Bank*, http://datatopics.worldbank.org/universal-health-coverage/covid19/, Accessed: 2020-04-16.

[3] T. Hale, S. Webster, A. Petherick, T. Phillips, and B. Kira, *Oxford COVID-19 Government Response Tracker*, https://github.com/OxCGRT/covid-policy-tracker/, Accessed: 2020-05-04, 2020.

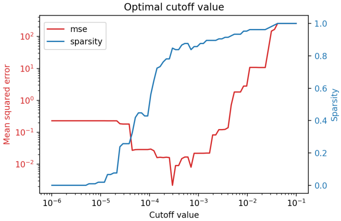
Appendix

Datasets and variables

Here is a more detailed look at the data we used:

- General COVID numbers for different countries from the John Hopkins University. Number of infected and recovered cases, and number of deaths by country/region.
- Information about government measures and restrictions provided by a group of researchers from Oxford [1]. Government response measures include school closing, workplace closing, travel bans and alike, an aggregation of these indicators into a "Stringency" index is also provided.
- Country indicators relevant in the context of the pandemic from the World Bank [2] (general health, hygiene and demographic indicators).

Sparsity and parametrization

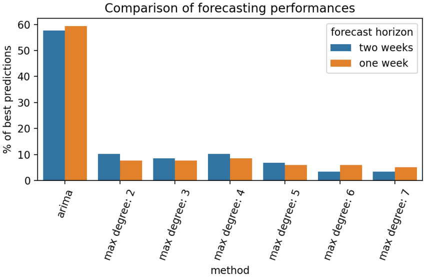


The choice of the cutoff value has a big impact on the resulting model:

- In a setting where all the required candidate functions are available to identify the dynamics the optimal model is quite sparse like in the above figure.
- When working with real data, more complex models tend to be the most precise ones.

In practice, we identify models with a wide range of cutoff values and plot their errors so that we choose an optimal cutoff value.

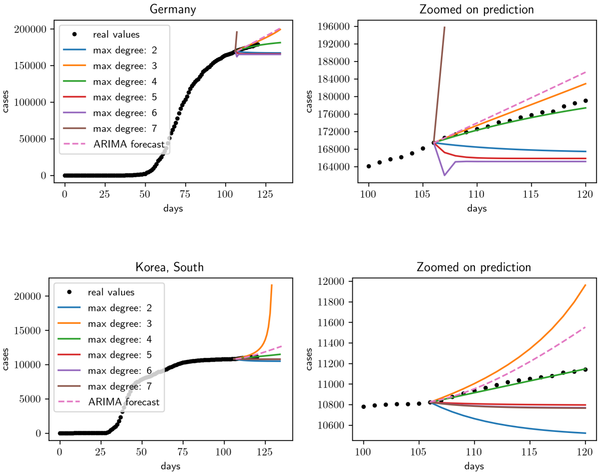
Forecasting capabilities



As depicted in the above figure, in more than half of the countries the statistical forecasting model showed better performance for both one week and two weeks forecast horizon.

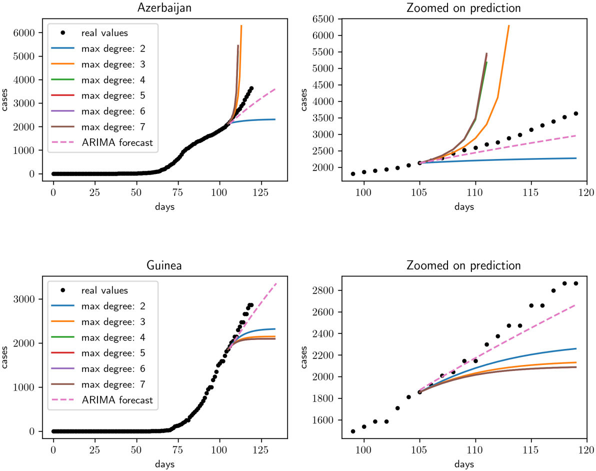
Forecasting capabilities

Successful examples:



Forecasting capabilities

Less successful examples:



Azerbaijan

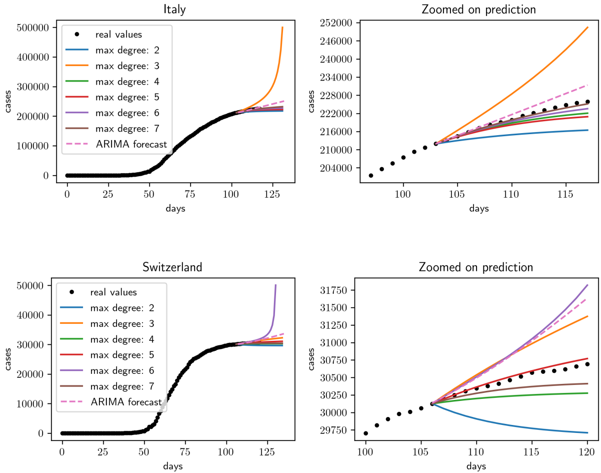
Zoomed on prediction

Guinea

Zoomed on prediction

Forecasting capabilities

Other examples:



More common epidemiological modeling techniques

The most common type of epidemiological models are the compartmental models introduced by [3] and based on compartments of the population like susceptible, exposed, infected, recovered. The most basic one is the SIR model:

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

Other approaches for modeling the spread of an infectious disease include:

- Spatiotemporal SIRs
- Statistical models
- Gravity models
- Network-based models
- Agent-based models

More thorough review of the different modeling approaches can be found in [4].

What state to choose?

In our work we only considered tracking the cumulative number of cases as the state **x** but as compartmental models suggest, it might be necessary to track the number of susceptible, infected, recovered people in order to understand the dynamics.

But when identifying the dynamics from simple generated SIR trajectories:

- We are not able to identify the original ODEs
- Identified dynamics do not fit data perfectly
- Resulting trajectories often diverge quite quickly

For real trajectories, results are not better.

Conclusion

Identification of dynamics from SIR quantities needs more investigation.

Appendix references I

[1] T. Hale, S. Webster, A. Petherick, T. Phillips, and B. Kira, *Oxford COVID-19 Government Response Tracker*, https://github.com/OxCGRT/covid-policy-tracker/, Accessed: 2020-05-04, 2020.

[2] *Understanding the Coronavirus (COVID-19) pandemic through data. World Bank*, http://datatopics.worldbank.org/universal-health-coverage/covid19/, Accessed: 2020-04-16.

[3] W. O. Kermack, A. G. McKendrick, and G. T. Walker, “A contribution to the mathematical theory of epidemics,” *Proc. R. Soc. Lond.*, vol. 12, 700â€“721, A115 1997. DOI: 10.1098/rspa.1927.0118.

[4] D. Chen, “Modeling the spread of infectious diseases: A review,” in. Dec. 2014, pp. 19–42, ISBN: 978-1-118-62993-2. DOI: 10.1002/9781118630013.ch2.