Extracting the governing equations of epidemics through sparse regression

L. Lober[1], and F.A. Rodrigues[1]
[1]University of São Paulo, São Carlos, Brazil.

Epidemic modeling is a growing field largely based on real-world phenomena, and with recent contributions to its methodologies given by network science. Determining the behavior of a given population in the onset of an epidemic is essential to plan policies to mitigate its effects on affected individuals, shorten the time of exposure to a given disease and direct resources to the most affected areas, with the recent COVID-19 pandemic as an well documented example.

In this work, our proposal is to extract the governing equations describing the time evolution of a given population where an epidemic spreading occurs, and for that end we employ a sparse regression model known as SINDy [1], or *Sparse Identification of Nonlinear Dynamical Systems*. The equations resulting from such approach can then be used to make predictions about the future states of that group. The main advantage of SINDy is its interpretability, given its capabilities of recovering the equations from the system's dynamics; and through them one can also acquire more information on the properties of such system.

The model functions by solving an approximation problem of the form

$$\dot{X} \approx \Theta(X)\Xi, \tag{1}$$

where X is the data on the epidemics in matrix form, $\Theta(X)$ forms the chosen set of basis functions to apply and Ξ will be the set of sparse coefficient vectors obtained through regression. We can then solve this system of equations to obtain predictions for a given time interval, or extract properties of interest from them to characterize the epidemic.

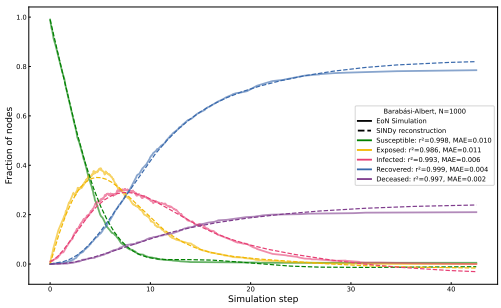

Fig. 1. Simulation of an SEIRD epidemic dynamics on a Barabási-Albert network with a thousand nodes. The resulting of applying the sparse regression model is shown on dashed lines with accuracy metrics for each curve.

Figure 1 shows an application of this method to a synthetic dataset generated by simulating a SEIRD, or an epidemic using the number of susceptible, exposed, infected, recovered and deceased cases as features, on a Barabási-Albert network with the EoN package [2]. We have also applied this model to several other epidemic models and network topologies with similar accuracy in the predictions, which will be thoroughly discussed in the presentation session of this work.

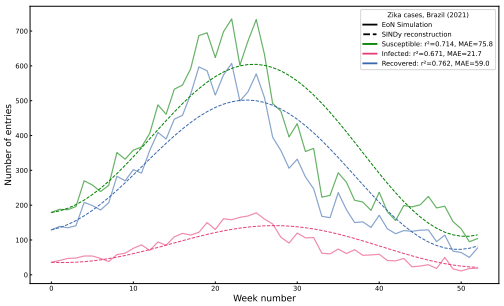

Fig. 2. Sparse regression reconstruction (dashed lines) on data of Zika virus in Brazil in the year of 2021 [3].

We have also begun exploring the capabilities of such model when using real epidemic data, as shown in Fig. 2, where SINDy is used to model data on the cases of Zika virus contagions in Brazil for the year of 2021. It can be seen that the model is capable of identifying the pattern in the evolution of each category used to classify the population exposed to this disease, even with the non-smooth behavior that is expected from real data.

[1] S. L. Brunton, J. L. Proctor, and J. N. Kutz. *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*. PNAS (2016).

[2] Miller et al., (2019). *EoN (Epidemics on Networks): a fast, flexible Python package for simulation, analytic approximation, and analysis of epidemics on networks*. Journal of Open Source Software (2019)

[3] Ministério da Saúde, *DATASUS*, https://datasus.saude.gov.br.