Discover governing reactions from concentration data

October 27, 2017

Introduction

When presented with a time series of possibly noisy concentrations as output of, e.g., measurements or simulations that were parameterized by microscopic rates (cite ReaDDy?), one can ask for the corresponding macroscopic rates and generating reaction network. In this paper we present an application of the shallow learning method SINDy [1] which is able to identify generating parsimonious nonlinear dynamics in data that stems from dynamical systems. In our application we, opposed to the original method, do not only look for macroscopic rates of net species change but investigate the specific reactions that might have lead to the observations. We demonstrate the algorithm on a toy problem demonstrating the identification abilities.

The method

The underlying model that we want to fit the data to is a law of mass action type dynamical system. To this end, let S be the number of species, then the concentration data at a time t can be represented by a vector

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_S(t) \end{pmatrix} \in \mathbb{R}^S.$$

Further, one can choose R possible ansatz reactions with their respective reaction function

$$\mathbf{y}_r(\mathbf{x}(t)) = \begin{pmatrix} y_{r,1}(\mathbf{x}(t)) \\ \vdots \\ y_{r,S}(\mathbf{x}(t)) \end{pmatrix}$$

so that the change of concentration for species i at time t_i, is represented by the dynamical system

$$\dot{\mathbf{x}}_i(t) = \sum_{r=1}^{R} y_{r,i}(\mathbf{x}(t))\xi_r, \quad i = 1, \dots, S,$$

where ξ_r are the to-be estimated macroscopic rates.

When presented with a time series with T observations, the data can be represented as a matrix

$$\mathbf{X} = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_S(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_S(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_T) & x_2(t_T) & \cdots & x_S(t_T) \end{pmatrix} \in \mathbb{R}^{T \times S}.$$

Given this matrix, one can propose a library $\Theta(\mathbf{X}) = \begin{pmatrix} \theta_1(\mathbf{X}) & \theta_2(\mathbf{X}) & \cdots & \theta_R(\mathbf{X}) \end{pmatrix}$ of R ansatz reactions with corresponding reaction functions

$$\theta_r(\mathbf{X}) = \begin{pmatrix} \mathbf{y}_r(\mathbf{X}_1)^T \\ \vdots \\ \mathbf{y}_r(\mathbf{X}_T)^T \end{pmatrix} \in \mathbb{R}^{T \times S}, \quad r = 1, \dots, R,$$

where $\mathbf{X}_i$ denotes the i-th row in X. Applying the concentration trajectory to the library yields a data tensor $\Theta(\mathbf{X}) \in \mathbb{R}^{T \times S \times R}$. Following the approach of SINDy, the goal is now to find coefficients $\Xi = \begin{pmatrix} \xi_1 & \xi_2 & \cdots & \xi_R \end{pmatrix}^T$, so that

$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi = \sum_{r=1}^{R} \theta_r(\mathbf{X})\xi_r.$$

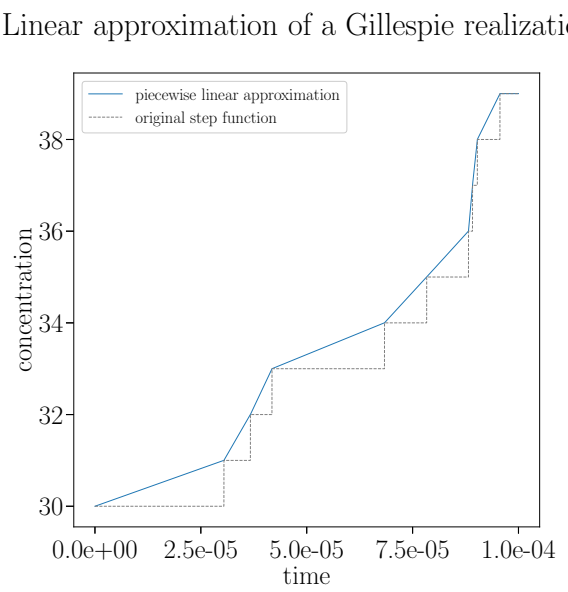

Figure 1: This figure depicts a possible output yielded by an application of the Gillespie method and the performed piecewise linear approximation in order to obtain a better behaving derivative.

In particular, the system is linear in the coefficients Ξ, which makes sparse regression tools such as LASSO [6, 3] applicable. To this end, one can consider the minimization problem to find $\hat{\Xi}$ such that

$$\hat{\Xi} = \arg\min_{\Xi} \left(\frac{1}{2T} \left\| \dot{\mathbf{X}} - \Theta(\mathbf{X})\Xi \right\|_F^2 + \alpha \|\Xi\|_1 \right) \quad \text{subject to } \Xi \geq 0, \tag{1}$$

where $\alpha \geq 0$ is a sparsity inducing hyperparameter. For $\alpha = 0$ this problem reduces to constrained least-squares. For solving (1) we apply the sequential least squares minimizer SLSQP, originally described in [5], contained in, e.g., the software package SciPy [4].

Examples

For generating time series data of concentrations we use the Gillespie method [2]. For a given set of initial conditions, we produce several realizations which are then converted to a trajectory with fixed time step and averaged. Since these trajectories are piecewise constant, one has to take special care when approximating the temporal derivative. Simply taking finite differences has the effect that, with decreasing time step, the derivative is most of the time constant zero and approaches infinity at the jump discontinuities. To counter this effect, we perform a linear approximation between each two adjacent discontinuities, as depicted in Figure 1.

Regression without regularization

Regression with regularization

Conclusion

dfdf

References

[1] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 1(609):1–26, 2015.

[2] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, New York, NY, 2009.

[4] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed October 27, 2017].

[5] D. Kraft. A software package for sequential quadratic programming. *Technical Report DFVLR-FB 88-28, Institut für Dynamik der Flugsysteme, Oberpfaffenhofen*, 1988.

[6] R. Tibshirani. Regression Selection and Shrinkage via the Lasso, 1996.