CEMRACS 2023 mini-project proposal

Title: Estimation of interactions in microbial communities *via* a neural network-based generalized smoothing algorithm.

Contact: Nicolas Brunel[(1)] (nbrunel@quantmetry.com), François Deslandes[(2)] (francois.deslandes@inrae.fr), Béatrice Laroche[(2)] (beatrice.laroche@inrae.fr), Lorenzo Sala[(2)] (lorenzo.sala@inrae.fr)
Institution or company: [(1)]Quantmetry Paris – ENSIIE, LaMME, Université d'Evry, [(2)]MaIAGE, INRAE

1. Context

The gut microbiota is a diverse collection of hundreds of microorganisms that play vital roles in digestion, metabolism, immune function, and neurological processes. Imbalances in this ecosystem have been associated with autoimmune and inflammatory diseases. Additionally, the gut microbiota serves as a protective barrier against pathogen invasion from ingested food. Recent advancements in sequencing techniques enable the identification of bacterial species and their abundance in fecal samples. The objective is to comprehend the interactions between these bacteria, their relationship with pathogens, and their functions within the ecosystem. A common approach in literature is to describe these interactions *via* the Generalized Lotka-Volterra (GLV) model [1]:

$$\frac{1}{x_i}\frac{\partial x_i}{\partial t} = \mu_i + \sum_{j=1}^{N} a_{ij} x_j \qquad (1)$$

where $x\_i(t)$ is the quantity of bacteria $i \in [1, \dots, N]$ at time t, μ_i defines the intrinsic growth rate of the population of bacteria i, and a_{ij} are the coefficients that represent the interactions between bacteria i and bacteria j. In the sequel m will denote the vector of N intrinsic growth rates and A the interaction coefficient matrix.

One of the main challenges in this context is that the bacterial data presents a number of samples significantly lower than the species of bacteria (low number of individuals for which a low number of time points are sampled). It should also be noted that direct estimation of the GLV model parameters (such as Maximum Likelihood estimation, Bayesian estimation, or even genetic algorithms) is not easy since it involves simulating the model for a wide range of parameters while exploring the high dimension parameter space, even though the latter is known to easily become stiff or even chaotic (existence of local minima, instability of the system in certain parameter regions). Moreover, these methods do not take advantage of the linearity of (1) w.r.t the model parameters. This is why in previous work we turned to an alternative method proposed by several authors, among which Ramsay and coworkers [2], [3] that we adapted in [4]. The proposed method uses splines in order to represent the abundances of bacterial species across time, which should be close to the experimental data while being also solution of the GLV model with unknown parameters m and A. Consequently, the proposed method performs a joint estimation of the spline coefficients and the model parameters through the alternate minimization of a goal function, which takes into account the proximity of splines to the data, a penalty related to the deviation with which the splines are a solution of a GLV model depending on the parameters to be estimated and a sparsity penalty on these parameters. In particular the algorithm is defined by the following steps:

- initialization – spline smoothing of data
- at each iteration
  - step 1 – regression for gradient matching (minimization of model parameters and sparsity penalty)
  - step 2 – data smoothing with model penalty (minimization of spline coefficients).

The advantage of using this alternate minimization is that (i) it can be easily implemented for large sized optimization problems, (ii) if the model is linear in the parameters step 1 is a convex quadratic optimization problem, thus it can be solved using highly efficient Nesterov accelerated proximal

gradient method, and (iii) if the system is linear also in the state, then it is a bi-convex optimization problem, where the convergence toward a stationary point is guaranteed.

2. Description and objectives

This project aims at improving the efficiency of the GS algorithm.
The main objective will be to employ a data-driven physics informed neural network (PINN) [5] to replace the spline smoothing and provide predictions of the abundances of bacterial species, since this step is the costly part in the estimation process. We therefore hope to speed up considerably the overall computational cost of the GLV parameter identification.
We will provide a Github repository with a code solving the GLV model using classical methods and a code performing the GS algorithm presented in [4].

3. Proposed methodology

We propose the following workflow:

1. introduction to the GLV model and familiarize with the classical GS algorithm;
2. develop a PINN to solve the GLV model with given parameters, investigating different neural networks architectures [6] (*e.g.* Multilayer Perceptrons);
3. develop a neural network to replace the splines in the GS algorithm and train it on a synthetic dataset (created using the PINN at previous step);
4. depending on the progression of previous steps, develop a neural network to replace the overall GS algorithm, both data trajectories and GLV model parameters estimation;
5. compare the efficiency and the accuracy of the new algorithm with respect to the GS algorithm on a real dataset.

4. Software requirements

- Python (numpy, scipy, pandas, matplotlib, scikit-learn, PyTorch)
- Github
- C/C++ compiler

5. References

[1] V. Volterra, “Lecons sur la Théorie Mathématique de la Lutte par la Vie,” *Gauthier-Villars Paris*, vol. 193, no. 1, 1931.

[2] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao, “Parameter estimation for differential equations: a generalized smoothing approach,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 69, no. 5, pp. 741–796, 2007, doi: 10.1111/j.1467-9868.2007.00610.x.

[3] A. A. Poyton, M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay, “Parameter estimation in continuous-time dynamic models using principal differential analysis,” *Comput. Chem. Eng.*, vol. 30, no. 4, pp. 698–708, Feb. 2006, doi: 10.1016/j.compchemeng.2005.11.008.

[4] N. Brunel, D. Goujot, S. Labarthe, and B. Laroche, “Parameter estimation for dynamical systems using an FDA approach,” in *11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)*, 2018.

[5] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *J. Comput. Phys.*, vol. 378, pp. 686–707, Feb. 2019, doi: 10.1016/j.jcp.2018.10.045.

[6] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, “Dive into Deep Learning.” arXiv, Feb. 10, 2023. doi: 10.48550/arXiv.2106.11342.