Predicting Microbial Community Interactions using Physics Informed Neural Networks.

Supervision

Lorenzo Sala, Beatrice Laroche from team Dynenvie, MaIAGE, INRAE.
Hugo Gangloff, Nicolas Jouvin from team SOLsTIS, MIA Paris-Saclay, INRAE.

Contact

Lorenzo Sala, `lorenzo.sala@inrae.fr`

To apply send an email with CV and motivation letter with subject 'APPLICATION FOR PINNs INTERNSHIP'.

Scientific context

The gut microbiota comprises a vast array of hundreds of microorganisms crucial for functions such as digestion, metabolism, immune response, and neurological processes. Disruptions in this complex system have been linked to autoimmune and inflammatory conditions. Moreover, the gut microbiota acts as a defense mechanism against the invasion of pathogens introduced through ingested food. Recent advances in sequencing technologies allow for the precise identification of bacterial species and their quantities in fecal samples. The goal of our work is to understand the relationships and interactions among these bacteria, their associations with pathogens, and their roles within the ecosystem. A common approach in literature is to describe these interactions via the Generalized Lotka-Volterra (GLV) model [5]:

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

where $x_i(t)$ is the quantity of the bacteria $i \in [1, \ldots, N]$ 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.

In this context, a significant challenge arises due to the bacterial data having a considerably lower number of samples compared to the multitude of bacterial species, with a small number of individuals sampled over a limited timeframe. It is important to highlight that directly estimating parameters for the GLV model—whether through Maximum Likelihood estimation, Bayesian estimation, or genetic algorithms—is challenging. This challenge stems from the need to simulate the model across a broad parameter spectrum, exploring a high-dimensional parameter space prone to stiffness and uncertainties, with the presence of local minima and system instability in certain parameter regions. Additionally, these methods fail to capitalize on the linearity of (1) concerning the model parameters. In previous works we turned to alternative methods, for instance the one proposed by Ramsay and coworkers [4, 2], adapted in [1]. 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. As a result, the proposed approach concurrently estimates spline coefficients and model parameters by iteratively minimizing an objective function. This objective function considers the proximity of splines to the data, a penalty associated with the deviation of the splines as a solution to the GLV model based on the parameters to be estimated, and a sparsity penalty on these parameters. The use of alternate minimization offers several advantages: (i) it is easily implementable for large optimization problems, (ii) if the model is linear in the parameters, the convex quadratic optimization problem is solvable with the highly efficient Nesterov accelerated proximal gradient method, and (iii) if the system is linear in the state as well, it becomes a bi-convex optimization problem with guaranteed convergence toward a stationary point.

In this context a substitute approach to the splines are the employment of Physics Informed Neural Networks (PINNs). We investigated the use of this hybrid machine-learning technique as a parametric approximation of the trajectories describing the abundances of bacterial species across time during the hackathon of CEMRACS 2023[1].

[1] `http://smai.emath.fr/cemracs/cemracs23/`

Internship objectives

This project aims first at improving the efficiency of the data-driven algorithm firstly developed during the CEMRACS to provide predictions of the abundances of bacterial species. The second goal is to develop, in the PINNs framework, an optimization procedure for the joint estimation of the GLV parameters and of the neural network parameters. To this goal, one could develop on existing approaches in the literature such as iterative methods [3] or direct estimation [6]. The numerical environment for the project is the Python library `jinns`[2] developed at MIA Paris-Saclay.

Activities assigned

1. Conduct, analyze, and present the state of the art on the topic.
2. Familiarize yourself with the Python library `jinns` and with the existing code.
3. Propose improvements and make modifications of the existing method.
4. Implement the alternative approach based purely on PINNs.
5. Compare the efficiency and the precision of the developed algorithms in 3 and 4 on synthetic and real data.
6. Analyze and communicate the results both orally and in writing.
7. Suggest future directions.

Internship overview

The candidate will benefit from the expertise of researchers of Maiage's modelling team, as well as biological insights and direct access to `jinns`' package developers.

The computing cluster of the unit is composed of 2 Dell PowerEdge R640 servers for CPU computing, 1 DELL Precision 5820 server for GPU test computing and the possibility to access the GPU cluster Lab-IA of the IDRIS Data Center.

- Profile: M2 students in applied mathematics or related fields.
- Duration: 6 months (flexible, ideally starting from April 2024).
- Placement location: The intern will be based at MAiAGE, at the INRAE center of Jouy-en-Josas (78). Regular meetings with the team SOLsTIS at MIA Paris-Saclay on the Campus Paris-Saclay.
- Project: This internship is funded by ANR PIA funding: ANR-20-IDEES-0002.
- Internship Allowance: $\sim$ 550 EUR / month.

Candidate background and skills

The candidate should have a background in modeling, dynamical systems and deep learning. Proficiency in Python scientific programming is required. We are seeking a candidate with a strong interest in working at an interdisciplinary level between mathematics and biology.

References

[1] 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.

[2] 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. *Computers & chemical engineering*, 30(4):698–708, 2006.

[3] 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. *Journal of Computational physics*, 378:686–707, 2019.

[4] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5):741–796, 2007.

[5] V. Volterra. Lecons sur la théorie mathématique de la lutte par la vie. *Gauthier-Villars, Paris*, 193(1), 1931.

[6] A. Yazdani, L. Lu, M. Raissi, and G. E. Karniadakis. Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLoS computational biology*, 16(11):e1007575, 2020.

[2] `https://gitlab.com/mia_jinns/jinns`