# Predictive modeling and parameter inference of gut microbial community from time-series data

Luana Ferreira

under the direction of
Dr. Kirill Korolev
and Dr. Feng Wang
Department of Physics
Boston University

Research Science Institute
July 28, 2015

## Abstract

The gut microbiota is a huge non-closed ecosystem. To study the interactions between it and the environment requires a lot of parameters. The generalized Lotka-Volterra (gLV) model can not explain the interaction with the environment. Thus, we tried to infer the immigration rate, now the gLV also written as $\dot{x_i} = m_i + x_i(r_i - \sum_j c_{ij}\ x_j)$, where $x$ is the abundance of species $i$ and $j$, $c_{ij}$ is the interaction between species and $r_i$ is the growth rate. Basically, we inferred these parameters from time-series data and compared with the original model, and we can see that the $m_i$ causes fluctuations in the growth rate $r_i$.



## Summary

The purpose of this research is to infer the immigration rate from time-series data in the generalized Lotka-Volterra model and to compare with the original model, do as to analyze interactions between the microbial community and the human gut. We also want to predict future evolutions of gut microbiome systems in various cases. For this purpose we use linear regression simulated in Python.

# 1 Introduction

The human body is a huge ecosystem of microbes. This is particularly true in the gastrointestinal tract of human, where most microbes live. Understanding the nature of interpopulation interactions in host associated microbial communities is crucial to understanding gut colonization, responses to perturbations, and transitions between health and disease [1]. Research on macroscopic ecosystems has been carried out for over a century, and many mathematical models could be potencially applied to study the gut microbial community. Most gut microbes are either harmless or of benefit to the host. The gut microbes protects against enteropathogens, extracts nutrients and energy from our diets, and contributes to normal immune function. The first step in understanding the symbiotic relationship between gut microbes and their host is to characterize the baseline healthy microbiota and the differences caused by disease. The gut microbiota is immensely diverse, varying between individuals and fluctuating over time, especially during times of disease and early development [2].

## 1.1 The Generalized Lotka-Volterra Model

Most of the techniques applied to model microbial communities are extremely useful for uncovering pair-wise interactions but do not adequately describe the intrinsic dynamic nature of them [1]. The generalized Lotka-Volterra model (gLV) simulates the interaction between species, generally the prey-predator dynamics. A classical example of the application of the Lotka-Volterra model is in the fox-rabbit-grass system, where the grass is the logistic term and it is part of the dynamics, and this is a closed system. The gLV also represents interactions amongst an arbitrary number of biological species. This interaction can be a predation or a competition, depending on the values of the parameters. Usually, a gLV model is written as:

$$\dot{x_i} = x_i(r_i - \sum_j c_{ij}\ x_j) \tag{1}$$

Where $x_i$ is abundance of species $i$, $r_i$ is the intrinsic growth rate, and $c_{ij}$ is the interaction from species $j$ to $i$.

The gLV model assumes pairwise taxon-taxon interaction. Usually, the interaction is inferred from covariance, which is a measure of how much two random variables change together, or co-occurence of the species abundances, resulting into a symmetric interaction due to the symmetry of covariance. The symmetric interaction is not generally true, since one species can inhabit another more than the reverse. So inferring gLV model is of particular importance to obtain an unbiased pairwise interaction. This also provides the parameters (immigration rate, growth rate, and pairwise interactions) needed to predict the future with the model. Due to the large number of species, one common way to do it is to reduce the number of species, usually using only 10-20 species for inference. There are more parameters than data, resulting in overfitting. But the problem is, even for previous research successfully inferred the parameters, those growth rates and pairwise interactions do not allow a stable coexistence of those species. Also the prediction by integrating the Ordinary Differential Equations (ODEs) is not always accurate. A successful inference means that the predict modeling is close to the experimental data. The gLV equations are non-linear and therefore allow for the existence of multiple steady states [3]. A missing part of the above model inference is immigration of microbes from environment, because the ecosystem is not closed, especially with the human gut. A gLV model with immigration could be written as:

$$\dot{x_i} = m_i + x_i(r_i - \sum_j c_{ij}\ x_j) \tag{2}$$

Where $m_i$ is the immigration rate.

## 1.2 The Gut Microbiota

Recent studies, propelled by metagenomics and next generation DNA sequencing technologies, have established novel connections between the intestinal microbial species composition and diseases. Recently, a mathematical model of microbiota dynamics that considers both species interaction networks and extrinsic perturbations such as antibiotics has been introduced. The model can explain how relatively simple ecological interactions such as competition for nutrients can lead to complex phenomena such as multi-stability or antibiotic-mediated catastrophic shifts. Most importantly, it was concluded that quantitative knowledge of the microbial interactions could enable the prediction of microbiota dynamics. The researchers extended the generalized Lotka-Volterra equations to infer microbiota ecology and predict its temporal dynamics under time-dependent external perturbations [3].

## 1.3 Competition Between Microorganisms

Interactions within ecological webs can have a positive impact (that is, a win), a negative impact (that is, a loss) or no impact on the species involved. A famous example of competition between microorganisms (that is, a loss-loss relationship) dates back from 1930s, which was executed a series of co-culture experiments. It was observed that for a number of species pairs, each species grew well in the absence of the other however, when co-cultured, one species (in this case, Paramecium aurelia) subdued the other (in this case, Paramecium caudatum). On the basis of these observations, the law of competitive exclusion was formulated on states that two species with similar niches exclude each other [4].

In this research we investigate the temporal development of microbes according to gLV, infer parameters of the gLV model with immigration (gLVi) from time-series data, and compare with the original gLV model to see if immigration can help to explain the time evolution of the gut microbial community.

# 2 Materials and Methods

For infering the parameters (immigration rate, growth rate, and pairwise interactions) in generalized Lotka-Volterra model with immigration (gLVi), (2) can be rewrite as:

$$\dot{x_i} = \frac{x_i(t + \triangle t) - x_i(t)}{\triangle t} = m_i + r_i \ x_i + \sum_j c_{ij} \ x_i \ x_j \tag{3}$$

The $x_i(t)$ data were obtained from Dataset2, an excel table found in Supporting Information of [1], where we chose 17 Operational Taxonomic Units (OTUs) to run the simulations. OTU is an operational definition of a species or group of species often used when only DNA sequence data is available, and it is the most commonly used microbial diversity unit. This Excel table provides the abundance of the OTUs in each day and each mouse of the experiment described in that paper. In this experiment, the data were collected by experimental practice mode, using mice, and the percentage of each type of bacteria was quantitavely collected. The researchers analyzed 27 kinds of bacteria in each mouse. To make the comparison across mice meaningful, we modeled the same OTUs across all the same five mice, as [1].

## 2.1 Linear Regression

Given the time-series data $x_i(t)$, all $x_i(t)x_j(t)$ are known, and the right side of the equation is linear with respect to the parameters $m$, $r$ and $c$, then linear regression can be applied to infer the parameters. The linear regression is a common technique, and can be done using LASSO, function from Scikit-learn, a Python package [5]. The linear regression may say if the prediction is close to data collected experimentally. We have the predicted model and the experimental data, and if the difference between the sample data and the predicted is small, the prediction works.

## 2.2 Infering Parameters

To perform linear regression in our case it is necessary to load mice data from the Excel table into Python as a matrix and then obtain 17 OTUs abundance and relative abundance for each day and each mice. The relative abundance can be obtained with matrix sum in Python. Thus, calculate $x_i(t+1) - x_i(t)$ for each OTU $i$ and day $t$ using matrix subtraction. After, calculate $x_i(t)$ and $x_i(t)\ x_j(t)$ for each OTU $i$ or OTU pairs $i$ and $j$ and day $t$. Then, we can do the linear regression using Scikit-learn:

$$x_i(t+1) - x_i(t) \text{ is the response variable;}$$
$$x_i,\ x_i(t)\ x_j(t) \text{ is the predictor variable.}$$

Finally we obtain the coefficients from the linear regression.

## 2.3 Softwares

To run the simulations, we also used a class called SciPy. This is a collection of mathematical algorithms and convenience functions built on the Numpy extension of Python [6]. From SciPy we also use Numpy and Matplotlib.pyplot as classes.

# 3 Results and Discussion

The parameter inference provided with graphs. The first two are a comparison between the gLV without and with the immigration rate. Both were plotted as true abundance, it means, the real data collected in the experiment with mice, versus the predicted abundance. In both cases, the predicted model can be consider acceptable due to the amount of points (real data) near the straight line. There is no significance difference between the model with and without immigration, in terms of abundance prediction, they have almost the same value of points predicted well. However, in the third plot we can see the prediction power of the gLV

over the time, with and without immigration. Once we have a $t_i$ abundance, we can predict the $t_{i+1}$ abundance. The distance between this two graphs measures how big is difference of prediction when the immigration rate is added. Thus, the generalized Lotka-Volterra model with immigration (gLVi) is more accurate than the gLV without immigration, because with immigration the quantity of days that can be predict is higher. It was predicted 5 days with the gLVi, while was predicted 3 days with the gLV. See the graphs in the following 3:

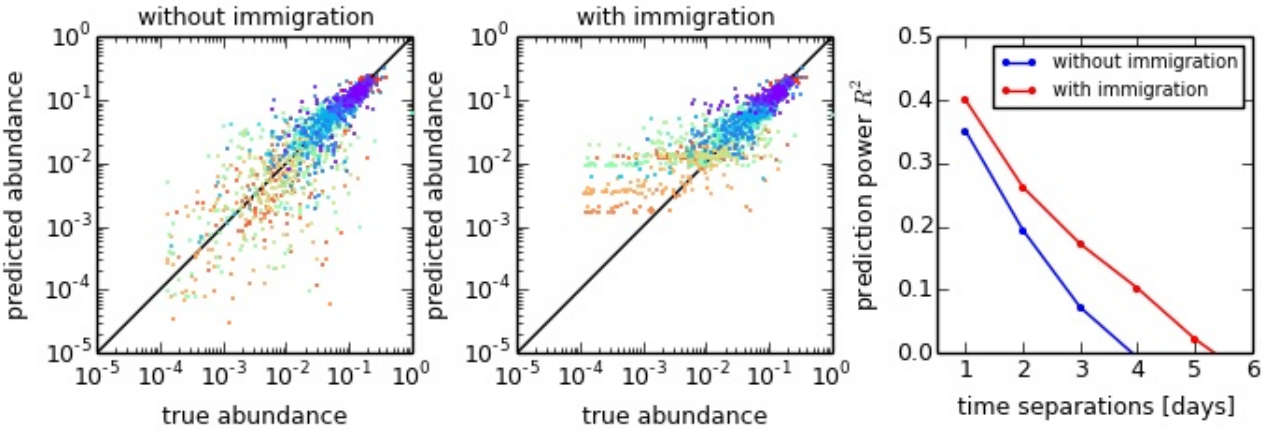


Figure 1: The first graph shows the simulation without immigration rate, and the second one shows the model with immigration rate; the third graph shows the prediction over the time.

## 3.1 Interactions between OTUs

Add the $m_i$ in the simulations could show that the gut microbiota is not a closed system, which is clear in the fluctuations of the $m_i$ parameter on graph upper left (see 3.1 below), there is an evident immigration from the human gut, showing another kind of interaction, with the environment. The graph alongside shows the $r_i$ (growth rate) in each OTU. It also has a fluctuation, but in the gLV with the $m_i$ the $r_i$ presents more negative growth rates than in the gLV without immigration rate. Then, death and branching dynamics is really

intense. The simulations also could provided with a OTU interaction graphs (see 3.1 below), where the two slower graphs represents the inferred interaction matrices of the 17 OTUs for the gLV and for the gLVi. These graphs shows that there is no much interaction between each other OTU when the immigration rate is considered. The kind of interaction is defined by the color, where blue is negative, white is null, and red is positive.

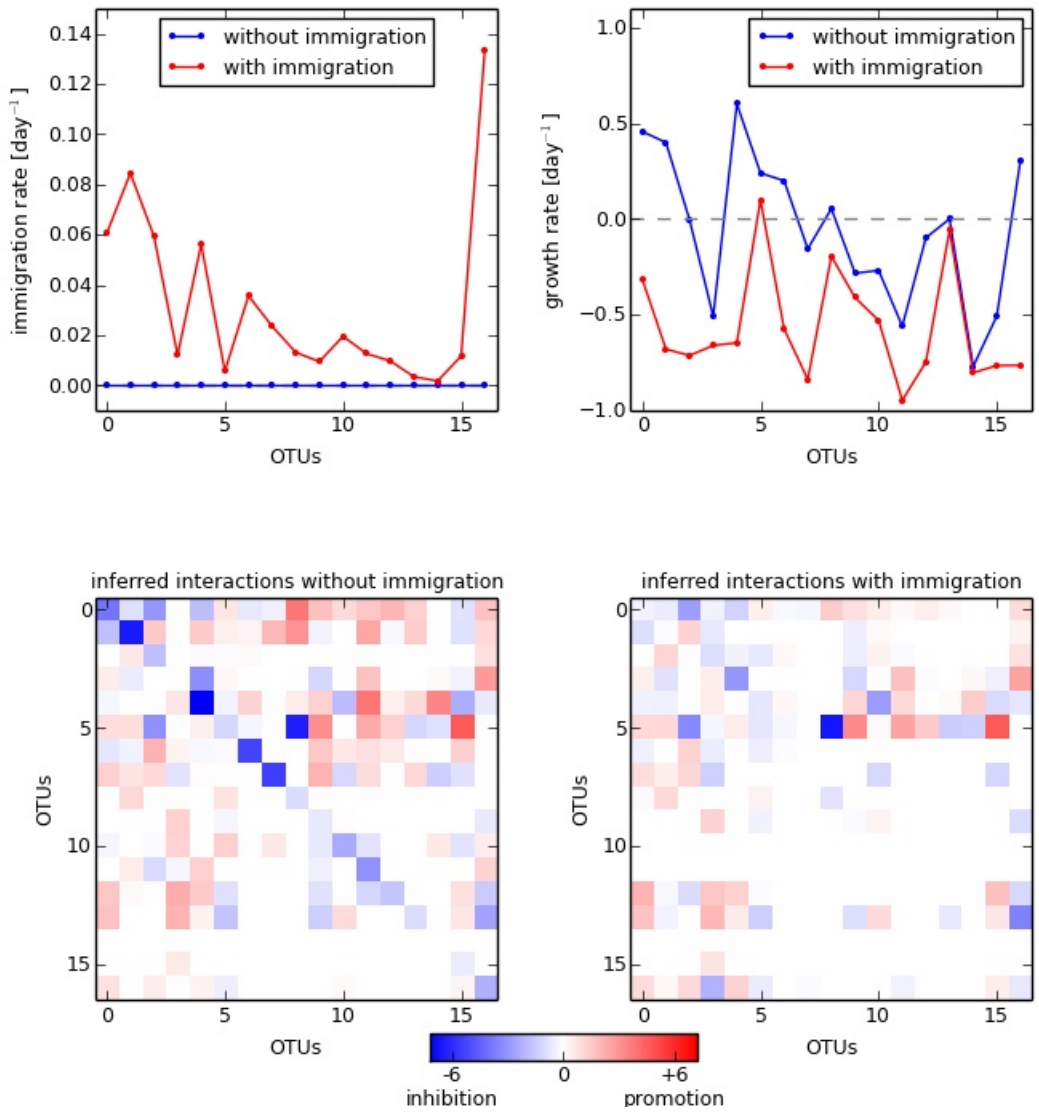


Figure 2: Graph upper left: $m_i$ vs. OTUs, graph upper right: $r_i$ vs. OTUs, graph slower left: interactions between OTUs without $m_i$, graph slower right: interactions between OTUs with $m_i$.

# 4 Conclusion

The comparison between gLV and gLVi showed that the immigration rate should be considered important when analyzing non-closed microbial systems, due to the interactions with the environment.

# 5 Acknowledgments

My mentors Dr. Kirill Korolev and Dr. Feng Wang from the Department of Physics at Boston University, for sharing their time, knowledge, and patience; My tutor Mr. Edward Vargas, for his comprehension and zeal; Dr. Jenny Sendova for helping me immensely and providing me with a great experience; I also would like to thank CEE (Center for Excellence in Education), MIT (Massachusetts institute of technology), and RSI (Research Science Institute) for this wonderful opportunity;

# References


[1] Simeone Marino, Nielson T. Baxter, Gary B. Huffnagle, Joseph F. Petrosino, and Patrick D. Schloss. Mathematical modeling of primary succession of murine intestinal microbiota. *Proceedings of the National Academy of Sciences of the United States of America*, 111(1):439–444, January 2014.

[2] Catherine A. Lozupone, Jesse I. Stombaugh, Jeffrey I. Gordon, and Janet K. Jansson Rob Knight. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489:220 – 230, 2012.

[3] Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar Rtsch, Eric G. Pamer, Chris Sander, and Joo B. Xavier. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*, 9(12):e1003388, 12 2013.

[4] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature*, 10:538 – 5500, 2012.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Introduction to scipy.org, 2008. Also avaiable as `http://www.scipy.org/`.