

SIMS: Social Inference of Microbiome Samples

By Tomer Oron & Yuval Dotan

i.d 322549312, 318505120

The research was performed in the School of Computer Science, Faculty of Exact Sciences

Under the supervision of Mr. Omri Peleg & Yadid Algavi

October 2024

## Introduction

The gut microbiome plays a major role in animal health and behavior1–3, influenced by individual factors like diet and genetics, as well as social interactions and environmental conditions4. Recent studies on social species have shown that individuals in the same social group share similar microbial communities5–7. In this project, we use data from the Amboseli baboon population8 to predict individual baboon microbiome composition by combining their own microbial profiles with those of their social group, based on data collected between 2000 and 20139.

## Methods

### Data Exploration

We used 6096 samples collected from 80 baboons for training and validation, based on 16S rRNA gene sequencing10, from fecal samples collected opportunistically. Each sample is includes metadata on environmental conditions and details about the baboon at the time of collection.

**The metadata** for each sample includes several components. Temporal information consists of the collection date, month, and hydrological year. At the individual level, the data specifies the baboon's unique identifier (*baboon\_id*), age at the time of collection, and sex. Social group data provides details on the baboon’s social group affiliation during sampling, the group's size, and diet composition, which is assessed using 30-day sliding windows and Principal Component Analysis (PCA)11The environmental context, covering the season (dry or wet) and the amount of rainfall during the collection month, is also included.Analysis of the training data revealed that each baboon has an average of 76.2 samples. The median time difference between consecutive samples from the same baboon is 22 days, with 282 samples having a preceding sample collected more than 180 days earlier**.** Although the wet season spans 42% of the year (November to May), 61% of the samples were collected during this period.

**For each sample**, there is a measurement of the abundance rates of the 61 most prevalent genera across all microbiome samples. Three analyses were conducted regarding samples from the same baboon, first, we compared the Bray-Curtis dissimilarity score between samples collected within the same date **(fig 1a)**, secondly, we calculated the Bray-Curtis dissimilarity score between a sample and the mean of the predeceasing samples **(fig 1b)**, and lastly, we calculated the score between every two concurrent samples **(fig 1c)**. These results suggest the predictive power of using the mean of previous samples, and using the last sample, while also providing us with the upper bound to the prediction accuracy one could expect.

A diagram of a graph

Description automatically generated with medium confidence

**Figure 1. Analysis results of the train data**

**(a)** Histogram of the time differences between the collection date of every two subsequent samples of the same baboon. **(b)** Kernel Density Estimation (KDE) smoothed histogram of the Bray- Curtis dissimilarity score between every two samples collected on the same date from the same baboon. **(c)** KDE smoothed histogram of the Bray- Curtis dissimilarity score between a sample and the mean of previous samples of the same baboon. **(d)** KDE smoothed histogram of the Bray- Curtis dissimilarity score between a sample and the previous sample of the same baboon.

### First Model

Building on our analysis, three key requirements were considered when constructing the model: 1. The model should use both the last sample and the mean of previous samples. 2. As the time difference between two consecutive samples increases, the influence of the last sample should decrease. 3. The model should present a seasonality effect.

Were is the sample to predict. The cosine function represents the seasonal effect, and the exponent reduces the effect of the last sample as the time difference increases. and are matrices representing the effect of each genus of bacteria. These parameters are learned for each baboon individually. is a scalar, a global parameter representing desegregation constant through time effect.

Using the training set, we trained the model's parameters through iterative optimization using the L-BFGS-B12 method. For all the baboons the optimization resulted in and . Thus, we decided that our first requirement should be removed, and we suggested a new model based on the average of previous samples. Also, the results of optimizing led us to the conclusion that the effect of bacteria genus on other genera is negligible, a conclusion raised by other studies13 as well.

### The Social Inference of Microbiome Samples (SIMS) Model

Based on the results of our initial model and studies showing that individuals within the same social group share similarities within their microbial composition, we refined the model to concentrate on circles of association: 1. The individual. 2. The individual’s immediate circle, its social group at the date of sampling. 3. The influence of broader circles, all the baboons which are not in its social group.

Were is the sample to predict at time t. is the average of samples taken within from , excluding those from this individual or baboons in its social group. is the average of samples taken within from , representing baboons in the individual's social group. is the weighted average of the previous samples, where . Here, and are vectors of size 61, representing the social inference – the effect of the individual, its social group, and the other baboons on its microbiome. is a scalar, a global parameter representing the decay in influence over time.

### Pipeline

In the preprocessing stage, the data is sorted by collection date, and the mean of samples from the same baboon on the same date is calculated. Subsequently, the data is split into sub-datasets, each corresponding to a specific baboon.

During the fitting stage, an initial grid search is conducted, followed by iterative optimization of β for each baboon and γ globally, utilizing the L-BFGS-B optimization method. The optimization process for each baboon involves using the mean of the Bray-Curtis dissimilarity function as a loss function. At the same time, the parameter γ is optimized by minimizing the sum of this loss across all trained baboons.

The prediction is calculated using configurable parameters. First, there is an option to choose between iterative and non-iterative prediction. In non-iterative prediction, the model predicts only based on the known samples, whereas in the iterative mode, the model also uses previously predicted samples. Another configurable parameter is the threshold, which specifies the minimum number of samples required to train beta for a new baboon. If this condition is not met, the model uses the average from the baboons already included in the model.

### Validation

We validated the results using cross-validation (4-fold). The training set was split into four random groups of size 20. We trained a model containing 60 baboons for each iteration and predicted the samples of the remaining 20 baboons in iterative and non-iterative modes. The validation was conducted on three scenarios: 1. Using very short time series (2 known samples per baboon), 2. Using short time series (10 known samples per baboon), and 3. Predicting ten latest samples given all other samples.

Code

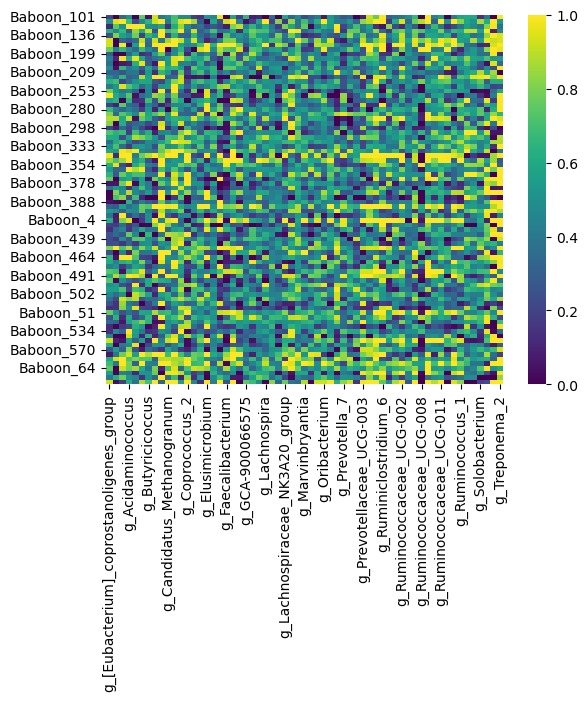
All code and analysis were written in Python, with NumPy, Matplotlib, SciPy, and Pandas. The source code is available at <https://github.com/yoavram-lab/Laland1995>.

## Results

While fitting the model using the complete training set, the optimal was found to be . The vectors have great variance between the baboons (fig 2), supporting studies that indicate the gut microbiome dynamics of baboons are highly individualized9.

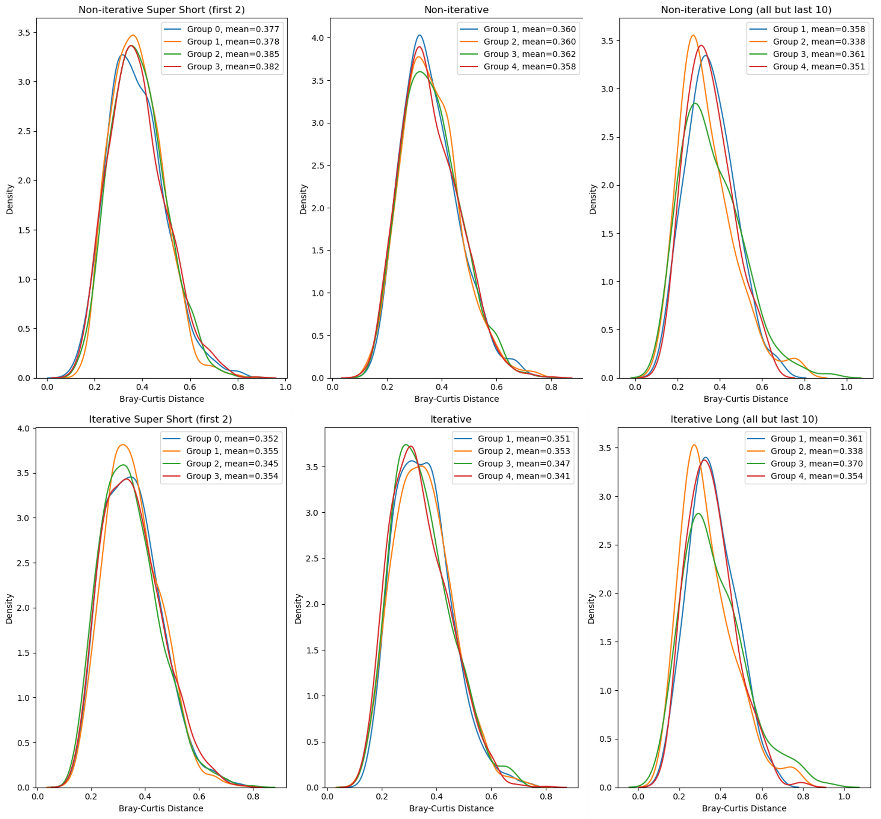
In scenario 1, the mean BC using an iterative mode was noticeably better than the non-iterative mode, whereas in scenario 3, the non-iterative prediction was preferable (fig 3). For scenario 2, we also checked the usage of trained per baboon in comparison to using the average of the other 60 baboons. The results showed for both cases that the iterative mode is slightly better for prediction and that using ten samples to train is not enough.

Our model gives better results than the naïve model that predicts using the mean of previous samples alone.



**Figure 2. Heatmap of the learned β vectors.**

Each row represents a baboon (y-axis) and each column corresponds to a microbial taxa (x-axis), labeled at the genus level. This visualization highlights the variation in β coefficients across different baboons and taxa.



**Figure 3. Cross validation results**

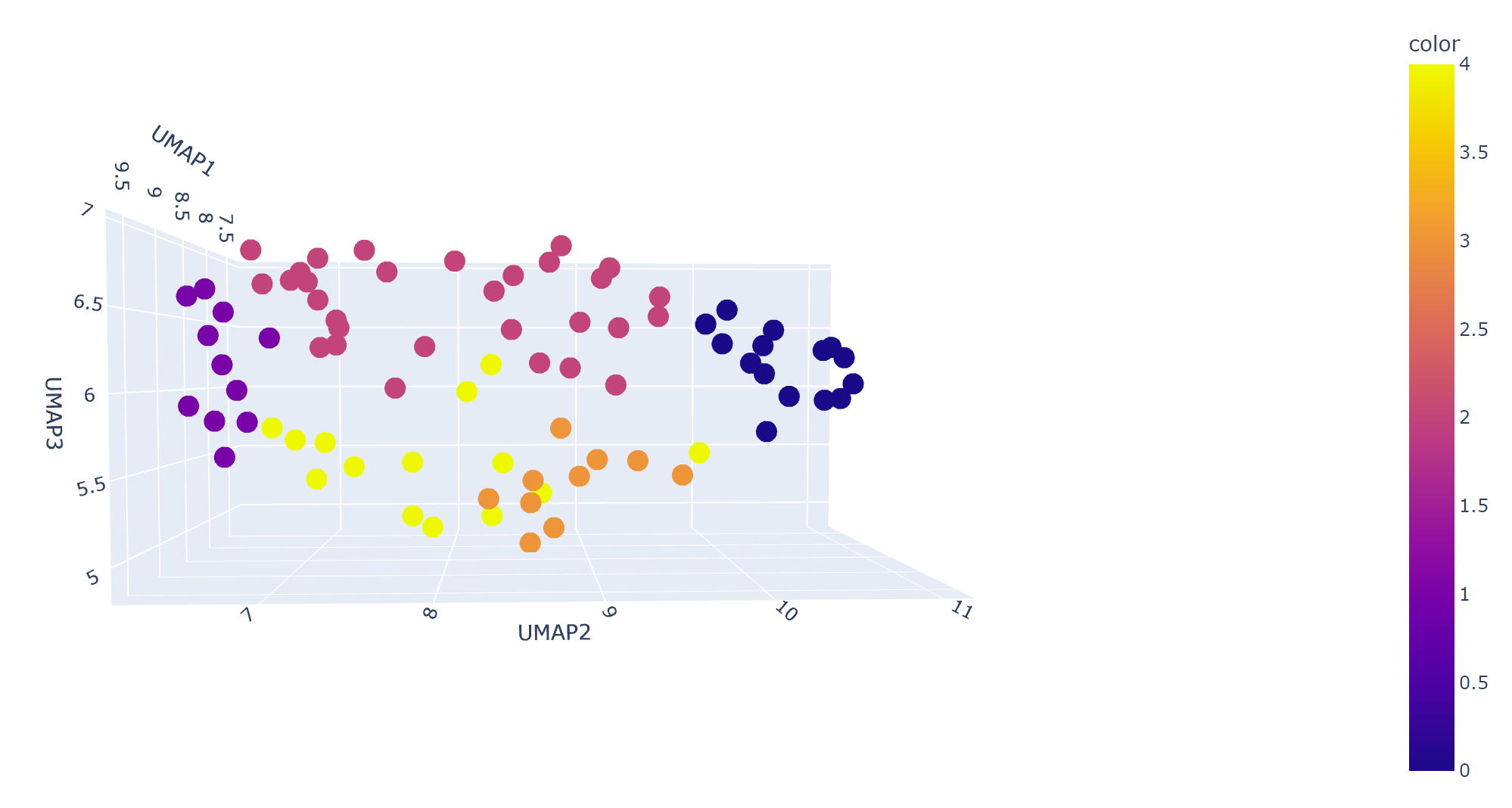
Kernel Density Estimation (KDE) plots of Bray-Curtis dissimilarity scores across different scenarios (short and long) and modes (non-iterative and iterative). Each line represents a test set of 20 baboons, with mean distances shown in the legend. The left column shows results for the first 2 samples, the middle for the first 10 samples, and the right for all but the last 10 samples.

## Discussion

During the development of our model, we questioned how to handle temporal data, specifically the seasonal effect and time gaps between samples. Throughout the model construction, we began with a simple model assigning equal weights to samples, then added the seasonal effect, and finally, we considered the time-decay factors. Using made the weights too small (effectively zero) due to the time differences between samples. Therefore, we decided to replace it with a function presenting slower decay, . In every model update, we have seen an increase in the prediction’s accuracy. Future work could explore other functions for seasonality and time decay to better capture temporal trends.

Another question we faced was whether using Newton-Raphson optimization algorithms was the correct method to find the best parameter. The Bray—Curtis dissimilarity function is defined using the L1 norm and, therefore, is not differentiable at every point. Throughout our work, we considered other optimization algorithms that do not require the function to be differentiable. Still, they did not present better parameters than the ones obtained through the L-BFGS-B algorithm and took more time to run.

While exploring the data, we used the K-means algorithm to cluster the baboons according to their average microbiome samples. We found 5 clusters (Fig 4) with a silhouette score of **X.** We did not find a correlation between the clusters and the social groups. These clusters may represent interactions between baboons other than social groups, which may improve the model. In order to do so, more metadata is required, such as family relations between baboons and more detailed data about each baboon's participation in social activities, such as grooming. Another method of finding relations between baboons is by clustering the trained betas of the baboons. The betas may also be used to find insights about correlations and interactions between bacteria.

[](umap_cluster5.html)

Bibliography

1. Iacob, S., Iacob, D. G. & Luminos, L. M. Intestinal Microbiota as a Host Defense Mechanism to Infectious Threats. *Front. Microbiol.* **9**, 3328 (2019).

2. Appleton, J. The Gut-Brain Axis: Influence of Microbiota on Mood and Mental Health. *Integr. Med. Clin. J.* **17**, 28–32 (2018).

3. Davenport, E. R. *et al.* The human microbiome in evolution. *BMC Biol.* **15**, 127 (2017).

4. Debray, R., Tung, J. & Archie, E. A. Ecology and Evolution of the Social Microbiome. *Annu. Rev. Ecol. Evol. Syst.* (2024) doi:10.1146/annurev-ecolsys-102622-030749.

5. Tung, J. *et al.* Social networks predict gut microbiome composition in wild baboons. *eLife* **4**, e05224.

6. Raulo, A. *et al.* Social networks strongly predict the gut microbiota of wild mice. *ISME J.* **15**, 2601–2613 (2021).

7. Wikberg, E. C., Christie, D., Sicotte, P. & Ting, N. Interactions between social groups of colobus monkeys (*Colobus vellerosus*) explain similarities in their gut microbiomes. *Anim. Behav.* **163**, 17–31 (2020).

8. Alberts, S. C. & Altmann, J. The Amboseli Baboon Research Project: 40 Years of Continuity and Change. in *Long-Term Field Studies of Primates* (eds. Kappeler, P. M. & Watts, D. P.) 261–287 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012). doi:10.1007/978-3-642-22514-7\_12.

9. Björk, J. R. *et al.* Synchrony and idiosyncrasy in the gut microbiome of wild baboons. *Nat. Ecol. Evol.* **6**, 955–964 (2022).

10. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).

11. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).

12. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization.

13. Roche, K. E. *et al.* Universal gut microbial relationships in the gut microbiome of wild baboons. *eLife* **12**, e83152 (2023).