

Logistic Tree Gaussian Processes (LoTGaP) for Microbiome Dynamics and Treatment Effects

by

Morris Greenberg

Department of Statistical Science
Duke University

Date: _____

Approved:

Li Ma, Advisor

Sayan Mukherjee

Pixu Shi

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

ABSTRACT

Logistic Tree Gaussian Processes (LoTGaP) for Microbiome
Dynamics and Treatment Effects

by

Morris Greenberg

Department of Statistical Science
Duke University

Date: _____

Approved:

Li Ma, Advisor

Sayan Mukherjee

Pixu Shi

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

Copyright © 2021 by Morris Greenberg
All rights reserved

Abstract

With advancements in and increased access to next-generation sequencing technology, hospitals (such as Duke Medical Center) have started to track the microbiomes of at-risk patients over time, but at inconsistently measured points across patients. Modeling the trajectories of high-throughput microbiome data proves difficult, due to inconsistent data collection, as well as a collection of analytical obstacles such as compositional data, sparsity, high dimensionality, and phylogenetic covariance structure. As a result, few methods allow us to capture uncertainty in the microbiome over time using increasingly standard data collection and processing methods.

Here, we develop a novel hierarchical model to measure dynamics of the microbiome across cohorts of patients measured inconsistently, which we call logistic-tree Gaussian processes for the microbiome (LoTGaP). LoTGaP adds to the existing microbiome literature through (1) using Gaussian processes to flexibly estimate the evolution of the microbiome over a finite set of days to handle missing/inconsistently measured data, (2) transforming operational taxonomic units (OTUs) to their internal nodes on the phylogenetic tree to accelerate computation and preserve biological relationships, and (3) building functionality to estimate the influence of covariates on microbiome dynamics across patients, which can allow for hospitals to link treatment regimens to microbiome dynamics, or make direct connections between microbiome data and other measurements, such as demographic information.

We demonstrate that LoTGaP produces uncertainty bands that reflect both within-person variation over time and across-person variation while comparing favorably in computation time to existing methods that are narrower in scope.

Acknowledgements

This research is partly supported by NIGMS grant 1R01GM135440.

I would like to thank Professor Li Ma for both his advisement throughout this work, and more generally his mentorship in my theoretical and technical development while at Duke. I would also like to thank the other members of my committee, Professor Sayan Mukherjee and Professor Pixu Shi, for their thoughtful critiques of my work. Additionally, I would like to thank Zhuoqun Wang and Dr. Pulong Ma, for their insights into the model development process, the other students in Li Ma's research group for their good questions in our group discussions, and Dr. Anthony Sung and team at the Duke Medical Center for providing the data used in this paper.

Special thanks to my parents for their continued support and encouragement in my studies, as well as my friends and colleagues at Duke who have only made my love for statistical science grow. In addition, thanks to Erica Voolich, Tamara Jenkins, Andy Andres, Fulton Gonzalez, and David Garman for being exemplary math educators in my middle school, high school, and undergraduate experiences. You helped shape my passion today.

Contents

| | |
|---|-------------|
| Abstract | iv |
| Acknowledgements | v |
| List of Figures | viii |
| List of Tables | ix |
| 1 Introduction | 1 |
| 1.1 Background | 1 |
| 1.2 Longitudinal Microbiome Studies | 1 |
| 2 Related Work | 3 |
| 2.1 Computational Considerations | 3 |
| 2.1.1 Compositionality | 3 |
| 2.1.2 Sparsity | 3 |
| 2.1.3 High-Dimensionality | 4 |
| 2.1.4 Phylogenetic Covariance Structure | 4 |
| 2.2 Probabilistic Time Series Modeling | 5 |
| 2.2.1 MALLARD | 5 |
| 2.2.2 TGP-CODA | 5 |
| 3 Methodology | 7 |
| 3.1 Hierarchical Model | 7 |
| 3.1.1 Modeling $y_{jt}(A_\ell)$ through latent $z_{jt}(A)$ | 7 |
| 3.1.2 Modeling $\mathbf{F}_j(A)$ through sampling $\mathbf{G}_j(A)$ from a Gaussian process and shifting by covariates matrix X_j | 10 |

| | | |
|----------|---|-----------|
| 3.2 | Posterior Estimation | 12 |
| 3.3 | Imputed Data Points | 13 |
| 4 | Results | 16 |
| 4.1 | Datasets | 16 |
| 4.2 | Analysis of Patients Sampled 4 Times in 4 Weeks | 16 |
| 4.3 | Analysis of Patients from Antiseptic Chlorhexidine Gluconate Study . | 19 |
| 4.4 | Run-Time Comparisons with TGP-CODA | 21 |
| 5 | Discussion | 23 |
| A | Pólya Gamma Augmentation | 25 |
| B | Full Conditionals | 26 |
| B.0.1 | Full conditionals on mean parameters based on Multivariate Gaussian Conditionals | 26 |
| B.0.2 | Full conditionals on variance terms based on Normal-Inverse Gamma Updates | 26 |
| B.0.3 | Full conditionals based on Pólya-Gamma augmentation (found in Appendix A) | 27 |
| | Bibliography | 28 |

List of Figures

| | | |
|-----|---|----|
| 3.1 | Example Phylogenetic Tree | 8 |
| 3.2 | Graphical Representation of LoTGaP | 12 |
| 4.1 | Trajectory of Lactobacillus and Enterococcus Internal Nodes | 17 |
| 4.2 | Distribution of Covariate Coefficients' Samples from LoTGaP run on CHG Study Samples | 20 |
| 4.3 | 2.5th vs. 97.5th Percentiles of Node-Covariate Samples | 21 |

List of Tables

| | | |
|-----|---|----|
| 4.1 | Node Split Descriptions in Figure 4.1 | 18 |
|-----|---|----|

List of Algorithms

| | | |
|---|--|----|
| 1 | Gibbs Sampler for node A | 13 |
| 2 | Gibbs Sampler for node A with imputation | 15 |

Chapter 1

Introduction

1.1 Background

Sequencing 16S ribosomal ribonucleic acid (rRNA) has been a gold standard to describe microbial communities for decades ([WBPL91], [WWB90]), since 16S rRNA (1) is pervasive in bacterial kingdoms and (2) contains highly variable regions across individual subjects along with conserved regions suitable for priming. With the advent of amplification of the 16S rRNA via polymerase chain reaction (PCR), biological researchers have been able to link clinical diagnoses of diseases ([HYJ⁺14], [SPS⁺12], [MSC⁺18]) with the microbial composition in cross-sectional studies.

Recently, there is movement towards associating diseases or other outcomes with the microbiome longitudinally, such as using microbial composition to predict mortality among allogeneic hematopoietic-cell transplantation (HCT) patients ([PGD⁺20]). With increasing evidence that there is no singular healthy microbiome ([Yon14]), tracking the trajectory of the microbiome within patients could offer increased statistical power to studies that previously only considered between-patient variation.

1.2 Longitudinal Microbiome Studies

At this time, human subject longitudinal microbiome studies are most commonly conducted by sampling at a few time points over a longer period. This often occurs in studies by sampling a baseline measurement and a follow-up measurement. For example, researchers have found that changes in habits such as diet ([AMN⁺18]) and exercise ([ZCS⁺18]) have implications on the microbiome. In both studies, the protocol involved taking measurements of two groups of patients, administering an intervention to one group (a high fiber diet or an exercise regimen) and no intervention to the other group, and recording a follow-up measurement after the treatment program concludes. There exist other study designs involving sampling points sparsely and consistently. For instance, a set of researchers tracked the development of the microbiome in newborns, taking measurements at 5 time points (2 days old, 3 months, 18 months, 3 years, and 5 years) ([HES⁺20]). While these are promising advances in our understanding of how the microbiome changes at a high level by making direct comparisons between static timepoints, these studies do not give any insight into how the microbiome evolves between scattered timepoints, and to what extent the observed changes are actual biological changes versus variation due to measurement error or daily fluctuation in the microbiome.

Researchers also employ artificial gut models to study the microbiome temporally ([MVWV93], [MMHV95], [BCL⁺20]). Artificial guts make it easy to collect samples at arbitrary frequencies in highly controlled settings. As a result, studies ranging from characterizing the behavior of oral drugs in children and adults ([BZB⁺04]) to describing the effect of prebiotics on metabolisms ([DPWF⁺10]) leverage *in vitro* methods of studying the microbiome. However, the sampling advantages from artificial gut models come at the cost of generalizability to the human microbiome. Differences between the human microbiome and artificial gut models are especially pronounced in immunocompromised patients such as colon cancer patients ([PCK⁺18]). It has been shown that the complexer the artificial gut model, the closer its microbiome is to human microbiomes ([CA19]), suggesting that artificial gut models may still have room for improvement in providing insight into the human microbiome.

There have been two studies involving frequent measurements of the microbiome. In [DMF⁺14], 2 healthy male subjects had their saliva and gut microbiota sampled daily over a year. In [CLC⁺11], one healthy male and one healthy female participant had their gut, saliva, and skin sampled daily over 15 months and 6 months, respectively. Both found that the microbiome is highly variable over time and can be linked with lifestyle habit changes such as diet, exercise, and sleep. While this setup may be ideal for observing the human gut microbiome, it is difficult at this point to enroll many subjects into a study of this nature because it is very time intensive. As a result, these studies sacrifice the ability to characterize between subject variation to be able to rigorously quantify within subject variation.

More recently, hybrid studies have started occurring, where multiple subjects are regularly measured (but not necessarily daily or at consistent time points across patients) over a short time period. These studies are the motivation behind the model proposed in this paper. The DIABImmune Microbiome Project ([Ins]) has tracked multiple samples of patients frequently over shorter time periods including studies analyzing the development of the microbiome in infants with Type-1 Diabetes ([KGS⁺15]), examining the differences in the microbiome from sets of families in 3 different countries to seek understanding of incidence rates of autoimmune diseases ([VKd⁺16]), and characterizing the microbiome in infants taking antibiotics ([YVS⁺16]). In addition, Memorial Sloan Kettering Cancer Center, in collaboration with Duke Medical Center, University Hospital Regensburg, and Hokkaido University Hospital, has collected over 10000 samples across over 1000 HCT patients used to analyze HCT microbially in multiple studies ([PGD⁺20], [STNL⁺19], [SPT⁺20]). Data collection from this project is ongoing, and samples of patients from Duke Medical Center from this project are used in this paper, discussed further in section 4.1.

Chapter 2

Related Work

2.1 Computational Considerations

We highlight key computational challenges associated with modeling the microbiome.

2.1.1 Compositionality

In PCR amplification, the number of reads assigned to specific operational taxonomic units (OTUs) or amplicon sequencing variants (ASVs) is constrained by an arbitrary constant sum that is conditional on the sequencer itself. As a result, absolute abundances of specific OTUs (derived from the random samples of molecules returned from PCR) have little to no meaning across patients ([GWPGE16]). A natural way to avoid considering the sequencing depth variation is to divide the absolute abundances by the total number of reads to yield proportions, or relative abundances, of the data. However, this transformation introduces a constraint where all of the relative abundances must sum to 1.

Compositional data analysis (CoDA) is commonly used to avoid this constraint ([Ait82]). In CoDA, variables are transformed using a generalized logistic transform so that they lie in a real coordinate vector space instead of the simplex. Alternatively, Dirichlet-multinomial models have been introduced as a way to model the counts directly ([HHQ12], [TC19], [CL13]). However, it has been suggested that Dirichlet-multinomial models impose too strict of assumptions on the covariance structure for the microbiome since they introduce negative correlations across all categories ([SL17]).

We discuss compositionality further in 3.1.1, where we propose an alternative method to avoid the simplex that incorporates biological information.

2.1.2 Sparsity

To obtain counts in microbiome data, it is common practice to match sequences from PCR output with a reference database of known OTUs. Above a viable threshold (historically, above 95% for genus matching, above 97% for species matching), instances of specific OTUs are recorded ([JSH⁺19]). There exists sparsity in OTU counts due to (1) matching algorithms' reliance on thresholds to record instances of OTUs (in which both false positive identifications of rare OTUs are sporadically introduced and false negative exclusions could occur for individual subjects), and (2)

heterogeneity in OTU composition across subjects (which causes individual samples to contain a small subset of the OTUs being tracked across all subjects).

Zero-inflated models and hurdle models have been introduced as means to handle excess zeros in the microbiome ([RBF⁺20], [HGZ18], [MAY18]). There is conflicting evidence of these models’ suitability for modeling of the microbiome. [XPTX15] finds that either zero-inflated or hurdle models are suitable, as long as the model choice is an extension of a distribution that can handle overdispersion (e.g., using a zero-inflated negative-binomial distribution over a zero-inflated Poisson distribution). However, [SRMD20] suggests avoiding zero-inflated models, as they can lead to spurious correlations in 16S rRNA data.

We directly address sparsity in the discussion section of this paper. We think it can most naturally be introduced with a multivariate extension of LoTGaP with sparse covariance structure across nodes. In addition, the logistic tree structure introduced in section 3.1.1 naturally introduces some sparsity at the taxa level, as seen in section 4.2, though can also cause computational instability when estimating across a strict subset of the subjects for which the logistic tree was built.

2.1.3 High-Dimensionality

Since the microbiome is heterogeneous across patients and consists of many unique microbes, microbiome data are high-dimensional and require modeling methods that handle high-dimensionality.

Dimension reduction techniques, such as Principle Coordinate Analysis (PCoA), are standards for both exploratory analysis and modeling of the microbiome ([Li15], [GDRP⁺14]). Often, researchers group OTUs to a higher taxonomic rank (such as species, family, or class) to naturally reduce dimension. There is also movement towards leveraging natural language processing techniques to categorize sets of OTUs based on overlap between patients ([TD20]) to reduce dimension.

We show in section 4.4 that LoTGaP’s parallelized and vectorized Gibbs sampling implementation allows for modeling high dimensional data without dimension reduction to be feasible.

2.1.4 Phylogenetic Covariance Structure

Abundances are estimated by matching sequences with known OTUs each defined by unique phylogeny. Consequentially, the covariance between two OTUs is conditional upon biological relationships.

Phylogenetic information is most commonly incorporated into microbiome modeling by transforming relative abundances to values that impose phylogenetic structure ([HLK10], [SWMD17]), and use these transformed data in models. Alternatively, penalized regression methods have been used to encourage taxa with close phylogenetic relationships to have similar coefficients ([TBJ14], [XCJ⁺18]).

The logistic tree method introduced in section 3.1.1 is another method that factors in phylogenetic information by modeling phylogenetic splits themselves rather than leaves on a phylogenetic tree. Additionally, in section 5, we discuss future extensions to further incorporate phylogenetic information, by imposing multivariate covariance structures based on phylogenetic distance.

2.2 Probabilistic Time Series Modeling

The impetus of this study is to build methodology for microbiome data taken at inconsistent time points across patients, which is scarcely considered in current literature, especially models that thoughtfully address most of the issues outlined in 2.1.1- 2.1.4. We highlight 2 recent microbiome time series developments below.

2.2.1 MALLARD

MALLARD is a multinomial logistic-normal dynamic linear model which can analyze biological and technical variation in the microbiome in artificial gut models [SDB⁺18]. Using an inverse isometric log-ratio (ILR) transform on abundances, MALLARD avoids simplex constraints and builds methodology for applying a staple Bayesian time series method, dynamic linear models, to the microbiome of an individual subject through a hierarchical model.

This hierarchical framework provides intuition to time series modeling of the microbiome. However, this model cannot be directly applied to the problem of interest, longitudinal modeling across multiple subjects with inconsistently measured data.

The authors extend MALLARD to handle (1) multiple patients (by using Kronecker products with indicator identity matrices, to model R independent Kalman filters - one for each of the R subjects) and (2) missing data (by using the Kalman filter to fill values). For our application though, information sharing between patients and time-points is desirable, and the structural independence in these extensions may be too inflexible for temporal modeling across subjects with missing data.

2.2.2 TGP-CODA

TGP-CODA is a temporal Gaussian process model for compositional data analysis which can analyze the trajectory of a single subject’s microbiome ([AMB18]).

Similarly to MALLARD, TGP-CODA uses a hierarchical model involving a generalized log-transform to link abundances to a time series method that is mapped on a real vector coordinate space rather than the simplex, and apply a common time series model for real data on the transformed microbiome data. Specifically, TGP-CODA uses the inverse additive log-ratio (ALR) transform to transform the abundances, often called the Softmax function, and employs Gaussian processes for

the time series modeling. We discuss the Softmax function and Gaussian processes further in sections 3.1.1 and 3.1.2, respectively.

TGP-CODA does not have any extension to multiple subjects at this time, which makes it limited in the types of studies it can analyze. In addition, the hierarchical model used in TGP-CODA makes it difficult to marginalize individual parameters, meaning that scaling the algorithm could prove challenging computationally.

Chapter 3

Methodology

We first show the generative model behind LoTGaP, then describe the contribution of individual parts, and finally illustrate a graphical representation of the model in figure 3.2. Thereafter, we outline computation of the model.

3.1 Hierarchical Model

Let A represent an individual internal node on the phylogenetic tree \mathcal{T} (a rooted, full binary tree, which is visually illustrated in figure 3.1) for $A \in \mathcal{I}_{\mathcal{T}}$, j represent an individual patient for $j = 1, \dots, J$, and t represent an individual time point for $t \in \mathcal{T}_j$. LoTGaP can be written as the following hierarchical model:

$$y_{jt}(A_\ell) \mid \theta_{jt}(A), y_{jt}(A) \stackrel{\text{ind}}{\sim} \text{Binomial}(y_{jt}(A), \theta_{jt}(A)) \quad (3.1)$$

$$z_{jt}(A) \mid \theta_{jt}(A), y_{jt}(A) \stackrel{\text{ind}}{\sim} PG(y_{jt}(A), F_{jt}(A)) \quad (3.2)$$

$$\text{where } F_{jt}(A) = \log \frac{\theta_{jt}(A)}{1 - \theta_{jt}(A)} \quad (3.3)$$

$$\mathbf{F}_j^T(A) \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{G}_j^T(A) + X_j \beta_A, \sigma_A^2 I_{|\mathcal{T}_j|}) \quad (3.4)$$

$$\mathbf{G}_j^T(A) \stackrel{\text{ind}}{\sim} \mathcal{N}(0, K_A(\mathcal{T}_j, \mathcal{T}_j \mid \eta_A, \rho_A)) \quad (3.5)$$

$$\text{where } K_A(\mathcal{T}_j, \mathcal{T}_j) = \begin{pmatrix} k(t_1, t_1 \mid \eta_A, \rho_A) & \dots & k(t_1, t_{n_j} \mid \eta_A, \rho_A) \\ \vdots & \ddots & \vdots \\ k(t_{n_j}, t_1 \mid \eta_A, \rho_A) & \dots & k(t_{n_j}, t_{n_j} \mid \eta_A, \rho_A) \end{pmatrix}$$

$$k(t, t' \mid \eta_A, \rho_A) = \eta_A^2 \exp(-\rho_A^{-2}(t - t')^2) \quad (3.6)$$

$$\sigma_A^2 \mid \Theta_\sigma \sim IG(\alpha_\sigma, b_\sigma) \quad (3.7)$$

$$\eta_A^2 \mid \Theta_\eta \sim IG(\alpha_\eta, b_\eta) \quad (3.8)$$

$$\beta_A \mid \tau^2 \sim \mathcal{N}(0, \tau^2 I_p) \quad (3.9)$$

where $y_{jt}(A_\ell)$ represents the absolute abundance of OTUs that are descendents of the left branch of node A , $X_j(A)$ represents a $|\mathcal{T}_j| \times p$ covariates matrix.

3.1.1 Modeling $y_{jt}(A_\ell)$ through latent $z_{jt}(A)$

We start by highlighting the intuition behind the logistic tree (equations 3.1-3.3) .

As mentioned in section 2.1.1, since the sequencing depth affects the absolute abundances in microbiome measurements, analyzing relative abundances becomes the standard to analyze the microbiome across samples with potentially different sequencing depths. Since the sum of relative abundances across a set of OTUs must sum to 1, a sample \mathbf{p} of microbiome data containing relative abundances of d OTUs lies in the d -simplex, $\mathbb{S}^d = \{\mathbf{p} = (p_1, \dots, p_d)^T : p_i > 0 \ (i = 1, \dots, d), p_1 + \dots + p_d = 1\}$, rather than \mathbb{R}^n (where n can be equal to d , but not required).

A popular method ([Ait82]), sometimes called compositional data analysis (CoDA), is to use a generalized logistic transformation from \mathbb{S}^d to \mathbb{R}^n , such as the additive log-ratio transformation, with inverse from \mathbb{R}^{d-1} to \mathbb{S}^d called the softmax transformation,

$$\text{Softmax}(\mathbf{x}) = \begin{bmatrix} \frac{\exp(x^{(1)})}{1 + \sum_{i=1}^{d-1} \exp(x^{(i)})} \\ \vdots \\ \frac{\exp(x^{(d-1)})}{1 + \sum_{i=1}^{d-1} \exp(x^{(i)})} \\ \frac{1}{1 + \sum_{i=1}^{d-1} \exp(x^{(i)})} \end{bmatrix}$$

While the softmax transformation maps data in \mathbb{R}^{d-1} to \mathbb{S}^d as we would hope, it introduces a new challenge in that marginalization of any individual components of \mathbf{x} becomes difficult computationally (due to the sum in the denominator). From a Bayesian hierarchical modeling perspective, this would force computation to be expensive through a Metropolis-Hastings algorithm or a Hamiltonian Monte Carlo method instead of a Gibbs Sampler.

We instead leverage the biological nature of microbiome data and construct a phylogenetic tree on the compositional data, where the absolute abundance of an individual OTU is a terminal node (leaf) on a binary tree, and branches and internal nodes are formed through evolutionary similarity between the different OTUs. Below is an example phylogenetic tree with absolute abundance of 2400 for 4 OTUs with relative abundances of $\frac{1}{4}$, $\frac{1}{6}$, $\frac{5}{12}$, $\frac{1}{6}$, and 3 internal nodes ($\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$):

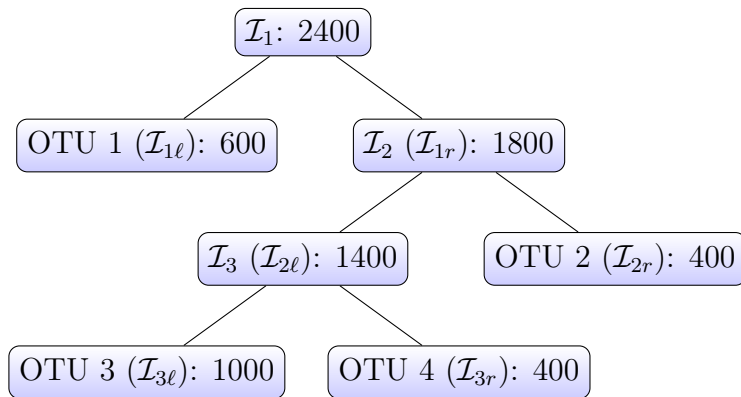


Figure 3.1: Example Phylogenetic Tree

At any individual internal node (a node where a split occurs) A of patient j 's phylogenetic tree \mathcal{T} at time t , we can model the absolute abundance of the left branch $y_{jt}(A_\ell)$ conditional on the absolute abundance at node A , $y_{jt}(A)$, as a binomial.

$$y_{jt}(A_\ell) \mid y_{jt}(A), \theta_{jt}(A) = \binom{y_{jt}(A)}{y_{jt}(A_\ell)} \theta_{jt}(A)^{y_{jt}(A_\ell)} (1 - \theta_{jt}(A))^{y_{jt}(A) - y_{jt}(A_\ell)} \quad (3.10)$$

$$\Rightarrow \mathcal{L}(\theta_{jt}(A); y_{jt}(A_\ell), y_{jt}(A)) \propto \theta_{jt}(A)^{y_{jt}(A_\ell)} (1 - \theta_{jt}(A))^{y_{jt}(A) - y_{jt}(A_\ell)} \quad (3.11)$$

where $\theta_{jt}(A)$ is the probability of going down the left branch at internal node A in the phylogenetic tree of patient j at time t , and \mathcal{L} is the likelihood function. Transforming $\theta_{jt}(A)$ to its logit, $F_{jt}(A) = \log \frac{\theta_{jt}(A)}{1 - \theta_{jt}(A)}$, we have the following extension to 3.11:

$$\mathcal{L}(F_{jt}(A); y_{jt}(A_\ell), y_{jt}(A)) \propto \frac{(e^{F_{jt}(A)})^{y_{jt}(A_\ell)}}{(1 + e^{F_{jt}(A)})^{y_{jt}(A)}} \quad (3.12)$$

This transformation still maps a simplex to a real vector space like the softmax, but can still allow for $F_{jt}(A)$ to be marginalized when introducing a latent Pólya-Gamma variable $z_{jt}(A)$. The Pólya-Gamma distribution is ideal here due to the following 2 integral identities ([PSW13]):

$$p(\omega \mid b, \psi) = \frac{e^{-\omega\psi^2/2} p(\omega)}{\int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega} \sim PG(b, \psi) \quad (3.13)$$

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega \quad (3.14)$$

where $\kappa = a - b/2$, $p(\omega)$ is the probability density function (PDF) of $PG(b, 0)$. Appendix A shows the full derivation of the Pólya-Gamma augmentation if we let $a = y_{jt}(A_\ell)$, $b = y_{jt}(A)$, $\psi = F_{jt}(A)$, $\omega = z_{jt}(A)$, based on equation 3.1.

At a high level, by introducing the latent variable $z_{jt}(A)$, we can express the conditional probability of $y_{jt}(A_\ell)$ in such a way that allows for conjugate computation for $F_{jt}(A)$, the log-odds of traversing down the left branch of internal node A in the phylogenetic tree for patient j at time t . Additionally, by modeling internal nodes and branches of the phylogenetic tree \mathcal{T} in this way, we avoid directly needing to deal with simplex constraints, while injecting biological characteristics of the microbiome to the generative model.

As a final note, the logistic tree structure has benefits in terms of (1) interpretation, and (2) sparsity compared to the isometric log-ratio (ILR), the generalized log-ratio transformation based on phylogenetic balance ([MSQ⁺17], [SWMD17]). By

the notation we have introduced for LoTGaP, the ILR can be represented as follows:

$$ILLR(A) = \sqrt{\frac{|A_\ell||A_r|}{|A_\ell| + |A_r|}} \log \left(\frac{g(y(A_\ell))}{g(y(A_r))} \right) \quad (3.15)$$

where $|A_\ell|, |A_r|$ represent the total number of leaves on the left and right subtrees, respectively, and $g(y(A_\ell)), g(y(A_r))$ are the geometric means of the abundances of leaves on the left and right subtrees, respectively. To understand why the ILR may be unintuitive, consider node \mathcal{I}_1 in figure 3.1. Its left subtree consists of 1 OTU with total abundance 600, while its right subtree consists of 3 OTUs with total abundances 400, 1000, and 400, resulting in $ILLR(\mathcal{I}_1) \approx \sqrt{\frac{1 \times 3}{1+3}} \log \left(\frac{600}{542.884} \right) = 0.0376$. By ILR, the tree is balanced towards the left subtrees at node \mathcal{I}_1 , despite the right subtree having three times the total abundance of the left subtree. On the other hand, the logistic tree reflects the balance in the tree by using sums instead of geometric means, meaning that by the logistic tree, \mathcal{I}_1 is balanced towards the right subtree in figure 3.1. The implications of using sums instead of geometric means on sparsity is also important. When a leaf has an abundance of 0, the geometric mean of an entire subtree containing that leaf is also 0, and the log-ratio of geometric means of a tree containing such a leaf is undefined. While past literature suggests adding arbitrarily small counts to such nodes or creating a zero-inflated ILR ([SWMD17]), it is much simpler to use the sums in the logistic tree framework as a remedy for sparsity.

3.1.2 Modeling $F_j(A)$ through sampling $G_j(A)$ from a Gaussian process and shifting by covariates matrix X_j

We now consider the temporal portion of the model (equations 3.4-3.9).

Research has shown that the gut microbiome can change within days ([DMC⁺14]). Simultaneously, taking microbiome measurements at hospitals among at-risk patients over time requires availability among both the hospital and the patient, to the point where gaps between measurements could be heterogeneous across patients.

Based on these factors, we model the microbiome temporally with Gaussian processes, which allows us to (1) have minimal assumptions on the functional form of the time series, and (2) easily adapt to inconsistently measured data. A Gaussian process is a collection of random variables, for which any finite subset of the random variables are jointly Gaussian. Notationally, a real stochastic process f is a Gaussian process over \mathcal{X} if and only if for any finite set $x_1, \dots, x_n \in \mathcal{X}$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \right) \quad (3.16)$$

where $m(x) = \mathbb{E}[f(x)]$, $k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$. In our case, we assume $|\mathcal{I}_{\mathcal{J}}|$ Gaussian processes, for which any patient j has had a finite set

of $|\mathcal{T}_j|$ realized observations from each process (where j 's finite subset of timepoints, $\mathcal{T}_j = t_1, \dots, t_{n_j} \in \mathcal{T}$). We use a squared exponential kernel, $k_A(t, t') = \eta_A^2 \exp(-\rho_A^{-2}(t-t')^2)$, $t, t' \in \mathcal{T}$, for the covariance so that measurements taken close to the data point in question have increased influence on the probability of a data point. In the squared exponential kernel, η_A^2 represents the signal variance (or in the context of the microbiome, the biological variance), and ρ_A represents how quickly the correlation between time points diminish as the gap increases. Note that in implementation of these models, we rescale the days such that if we perform inference on microbiome samples between days a and b , we set $a := 0, b := 1$ and $\forall x \in \mathbb{Z} \cap (a, b], x := \frac{x-a}{b-a}$.

In high-throughput DNA sequencing, technical variation exists across samples, and without replicates (as is the case in most human subject studies), it can be easy to overfit biological variation while underrepresenting technical variation. Therefore, we distinguish the finite drawing from a Gaussian process at internal node A for patient j , $\mathbf{G}_j(A)$, from the vector of the log-odds of traversing down the left branch of A at times $\{t_1, t_2, \dots, t_{n_j}\}$, $\mathbf{F}_j(A)$, through a noise term, which is distributed $\mathcal{N}(0, \sigma_A^2 I_{|\mathcal{T}_j|})$. We assume independence of the noise term across time points within a patient, as there is little to no evidence to suggest technical variation is correlated within a patient. However, other biases exist in 16S rRNA sequencing, such as batch effects ([LSB⁺10], [SGW11], [VBE⁺15]), the bias that occurs from sets of samples being processed under different conditions (such as laboratory conditions, personnel, etc.). For simplicity here, we do not estimate separate technical variations by batch, but acknowledge that LoTGaP can be extended to specify such generative processes.

Additionally, we allow for mean shifting of the Gaussian process at node A via linear basis functions on the set of covariates for patient j , X_j . Note that this framework allows for the user to extend the mean shifting to polynomial bases as long as the polynomials are explicitly input as unique covariates in the model. For computational ease, the coefficient vector β_A , is assumed to not change over time. In practical settings, constant coefficients over time enable us to (1) compare the microbiome between cohorts of patients (e.g., treatment group vs. control group in prospective studies, patients who contract a disease vs. those who do not in retrospective studies) and (2) condition on key demographic factors or baseline characteristics, such as age, gender, or body mass index (BMI) at the outset of tracking patients. While we eliminate the ability to relate the dynamics of the microbiome fluidly with another changing measure by not allowing for changing variables and coefficients over time, the inclusion of constant covariates empowers LoTGaP to perform inference on many relevant questions about the microbiome in hospital settings.

Equations 3.7-3.9 add conjugate priors to σ_A^2 , η_A^2 , and β_A . Other priors could be chosen, especially if variable selection is desired for β_A .

Below, we show a graphical representation of the model for an internal node $A \in \mathcal{I}_{\mathcal{T}}$, where grey, white, and green circles represent observed random variables, latent random variables introduced for inference, and latent random variables introduced for model fitting, respectively, and uncircled items are priors:

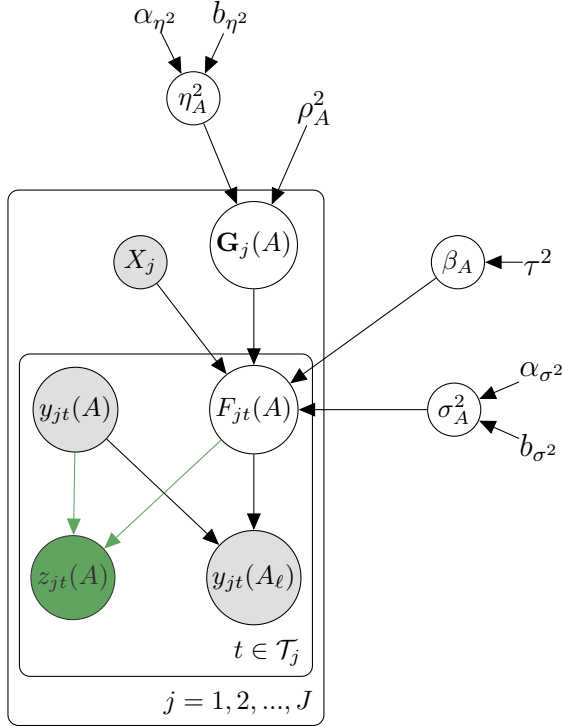


Figure 3.2: Graphical Representation of LoTGaP

3.2 Posterior Estimation

By using the Pólya-Gamma augmentation to model the logistic tree and using Gaussian processes to model time, all of the parameters in the model are conjugate and can be updated with a Gibbs sampling step except for ρ_A , the double exponential kernel decay rate in the Gaussian process. Multiple people have suggested priors for ρ_A for Hamiltonian Monte Carlo Gaussian process implementations ([Nea97], [Ras97]), but there has been little to no literature on conjugate priors in the context of a Gibbs sampler. At this point, we input ρ as a parameter that is constant across internal nodes A , with an eye towards adding a Metropolis-Hastings update step for ρ_A in the future.

Below is a Gibbs sampling algorithm for an internal node $A \in \mathcal{I}_{\mathcal{T}}$, based on the full conditionals derived in appendix B.

Algorithm 1: Gibbs Sampler for node A

```

for  $i = 1, \dots, K_{iter}$  do
  for  $j = 1, \dots, J$  do
    1. Update  $\mathbf{z}_j(A)$ :
    for  $t$  in  $\mathcal{T}_j$  do
       $z_{jt}(A)^{(i)} \leftarrow PG(y_{jt}(A)^{(i-1)}, F_{jt}(A)^{(i-1)})$ 
    end
    2. Update  $\mathbf{F}_j(A)$ 
     $\mathbf{F}_j^T(A)^{(i)} \leftarrow$ 
     $\mathcal{N}\left(\Sigma_{\mathbf{F}_j^{(i)}} \left[ \frac{1}{\sigma_A^{2(i-1)}} (\mathbf{G}_j^T(A)^{(i-1)} + X_j \beta^{(i-1)}) + \kappa_{jt}(A) \right], \Sigma_{\mathbf{F}_j^{(i)}}\right)$ 
    3. Update  $\mathbf{G}_j(A)$ 
     $\mathbf{G}_j^T(A)^{(i)} \leftarrow \mathcal{N}\left(\Sigma_{\mathbf{G}_j} \left[ \frac{1}{\sigma_A^{2(i)}} I_{|\mathcal{T}_j|} (\mathbf{F}_j^T(A)^{(i)} - X_j \beta_A^{(i-1)}) \right], \Sigma_{\mathbf{G}_j}\right)$ 
  end
  4. Update  $\sigma_A^2$ 
   $\sigma_A^{2(i)} \leftarrow IG\left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\sigma, \frac{1}{2} \sum_{j=1}^J D^{(i)T} D^{(i)} + b_\sigma\right)$ 
  where  $D^{(i)} = \mathbf{F}_j(A)^{(i)} - \mathbf{G}_j(A)^{(i)} - X_j \beta_A^{(i-1)}$ 
  5. Update  $\eta_A^2$ 
   $\eta_A^{2(i)} \leftarrow IG\left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\eta, \frac{1}{2} \sum_{j=1}^J (\mathbf{G}_j(A)^{(i)})^T K'(\mathcal{T}_j, \mathcal{T}_j)^{-1} \mathbf{G}_j(A)^{(i)} + b_\eta\right)$ 
  6. Update  $\beta_A$ 
   $\beta_A^{(i)} \leftarrow \mathcal{N}\left(\Sigma_{\beta_A^{(i)}} \left( \frac{1}{\sigma_A^{2(i)}} \sum_{j=1}^J X_j^T (\mathbf{F}_j^T(A)^{(i)} - \mathbf{G}_j^T(A)^{(i)}) \right), \Sigma_{\beta_A^{(i)}}\right)$ 
end

```

To speed up run-time on the Gibbs sampler of the entire phylogenetic tree, we parallelize the algorithm across all nodes on a tree and vectorize across \mathcal{T}_j .

3.3 Imputed Data Points

A key feature of Gaussian processes (and consequentially, LoTGaP) is being able to impute missing data points and provide uncertainty around them. In LoTGaP, for any patient j , under the Gaussian process with noise, we estimate $\mathbf{F}_j(A) \forall t \in \mathcal{T}_j$, whenever a measurement is taken in the set of \mathcal{T} days. Consider $\mathbf{F}_j^*(A)$, which consists of the points $t' \in \mathcal{T}_j^* = \mathcal{T} \setminus \mathcal{T}_j$, the set of days in the study where patient j did not have a measurement. We can write the joint distribution of $\mathbf{F}_j(A)$, $\mathbf{F}_j^*(A)$ as follows:

$$\begin{bmatrix} \mathbf{F}_j(A) \\ \mathbf{F}_j^*(A) \end{bmatrix} = \begin{bmatrix} \mathbf{G}_j(A) \\ \mathbf{G}_j^*(A) \end{bmatrix} + \begin{bmatrix} X_j \beta_A \\ X_j \beta_A \end{bmatrix} + \begin{bmatrix} \mathbf{e}_j(A) \\ \mathbf{e}_j^*(A) \end{bmatrix} \quad (3.17)$$

$$\sim \mathcal{N} \left(\begin{bmatrix} X_j \beta_A \\ X_j \beta_A \end{bmatrix}, \begin{bmatrix} K_A(\mathcal{T}_j, \mathcal{T}_j) + \sigma_A^2 I_{|T_j|} & K_A(\mathcal{T}_j, \mathcal{T}_j^*) \\ K_A(\mathcal{T}_j^*, \mathcal{T}_j) & K_A(\mathcal{T}_j^*, \mathcal{T}_j^*) + \sigma_A^2 I_{|T_j^*|} \end{bmatrix} \right) \quad (3.18)$$

Conditionalizing the above joint distribution, we have the following update for the imputed posterior predictive density at iteration i :

$$\mathbf{F}_j^*(A)^{(i)} \mid \mathbf{F}_j(A)^{(i)}, \beta_A^{(i)}, \sigma_A^{2(i)}, \tau^2, \mathcal{T}_j, \mathcal{T}_j^*, X_j \sim \mathcal{N} \left(\mu^{*(i)}, \Sigma^{*(i)} \right) \quad (3.19)$$

$$\mu^{*(i)} = X_j \beta_A^{(i)} + K_A(\mathcal{T}_j^*, \mathcal{T}_j \mid \eta_A^{(i)}, \rho_A) \times \quad (3.20)$$

$$(K_A(\mathcal{T}_j, \mathcal{T}_j \mid \eta_A^{(i)}, \rho_A) + \sigma_A^{2(i)} I_{|T_j|})^{-1} \times \\ \left(\mathbf{F}_j(A)^{(i)} - X_j \beta_A^{(i)} \right)$$

$$\Sigma^{*(i)} = K_A(\mathcal{T}_j^*, \mathcal{T}_j^* \mid \eta_A^{(i)}, \rho_A) + \sigma_A^{2(i)} I_{|T_j^*|} - \quad (3.21)$$

$$K_A(\mathcal{T}_j^*, \mathcal{T}_j \mid \eta_A^{(i)}, \rho_A) \times$$

$$(K_A(\mathcal{T}_j, \mathcal{T}_j \mid \eta_A^{(i)}, \rho_A) + \sigma_A^{2(i)} I_{|T_j|})^{-1} \times$$

$$K_A(\mathcal{T}_j, \mathcal{T}_j^* \mid \eta_A^{(i)}, \rho_A)$$

In algorithm form, we addend algorithm 1 by adding synthetic datasets of imputed points for any iteration after burn-in that is not thinned out:

Algorithm 2: Gibbs Sampler for node A with imputation

```

for  $i = 1, \dots, K_{iter}$  do
  Impute  $\leftarrow$  FALSE
  if  $(i - 1) > i_{burnin}$  and  $(i - 1) \bmod thin == 0$  then
    | Impute  $\leftarrow$  TRUE
  for  $j = 1, \dots, J$  do
    1. Update  $\mathbf{z}_j(A)$ :
    for  $t$  in  $\mathcal{T}_j$  do
      |  $z_{jt}(A)^{(i)} \leftarrow PG(y_{jt}(A)^{(i-1)}, F_{jt}(A)^{(i-1)})$ 
    end
    2. Draw imputed  $\mathbf{F}_j^*(A)$ 
    if Impute then
      |  $\mathbf{F}_j^{*T}(A)^{(i-1)} \leftarrow \mathcal{N}(\mu^{*(i-1)}, \Sigma^{*(i-1)})$ 
    3. Update  $\mathbf{F}_j(A)$ 
     $\mathbf{F}_j^T(A)^{(i)} \leftarrow$ 
     $\mathcal{N}\left(\Sigma_{\mathbf{F}_j^{(i)}} \left[ \frac{1}{\sigma_A^{2(i-1)}} (\mathbf{G}_j^T(A)^{(i-1)} + X_j \beta^{(i-1)}) + \kappa_{jt}(A) \right], \Sigma_{\mathbf{F}_j^{(i)}}\right)$ 
    4. Update  $\mathbf{G}_j(A)$ 
     $\mathbf{G}_j^T(A)^{(i)} \leftarrow \mathcal{N}\left(\Sigma_{\mathbf{G}_j} \left[ \frac{1}{\sigma_A^{2(i)}} I_{|\mathcal{T}_j|} (\mathbf{F}_j^T(A)^{(i)} - X_j \beta_A^{(i-1)}) \right], \Sigma_{\mathbf{G}_j}\right)$ 
  end
  5. Update  $\sigma_A^2$ 
   $\sigma_A^{2(i)} \leftarrow IG\left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\sigma, \frac{1}{2} \sum_{j=1}^J D^{(i)T} D^{(i)} + b_\sigma\right)$ 
  where  $D^{(i)} = \mathbf{F}_j(A)^{(i)} - \mathbf{G}_j(A)^{(i)} - X_j \beta_A^{(i-1)}$ 
  6. Update  $\eta_A^2$ 
   $\eta_A^{2(i)} \leftarrow IG\left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\eta, \frac{1}{2} \sum_{j=1}^J (\mathbf{G}_j(A)^{(i)})^T K'(\mathcal{T}_j, \mathcal{T}_j)^{-1} \mathbf{G}_j(A)^{(i)} + b_\eta\right)$ 
  7. Update  $\beta_A$ 
   $\beta_A^{(i)} \leftarrow \mathcal{N}\left(\Sigma_{\beta_A^{(i)}} \left( \frac{1}{\sigma_A^{2(i)}} \sum_{j=1}^J X_j^T (\mathbf{F}_j^T(A)^{(i)} - \mathbf{G}_j^T(A)^{(i)}) \right), \Sigma_{\beta_A^{(i)}}\right)$ 
end

```

Uncertainty is directly estimated from the synthetic draws from the Gibbs sampler. We show a descriptive example using the data imputation in section 4.2.

Chapter 4

Results

4.1 Datasets

16S rRNA data in sections 4.2-4.4 come from Duke Medical Center. Stool samples from patients were collected, stored, and processed as previously described in [GKR⁺21]. These data consist of 1575 samples, 793 of which are from patients with allogeneic HCT, 466 from patients with autologous HCT samples, and 316 samples from others, including healthy relatives of the transplant patients. There are 537 unique patients tracked.

Memorial Sloan Kettering Cancer Center processed sequence data via the `dada2`, `ShortRead`, `seqinr`, and `ape` packages, and used the 16S rRNA database from the National Center for Biotechnology Information (NCBI) [NCB18] to map ASVs to taxonomy assignment. We carried out further data cleaning via the `phyloseq`, `phangorn`, and `DECIPHER` packages in R to quality filter sequences, identify amplicon sequencing variants (ASVs), align sequences across samples, and create the phylogenetic tree. We applied a prevalence filtering threshold of 0.01 and mean relative abundance threshold of 0.0001, and removed any ASVs with no phylum found. The resulting phylogenetic tree was constructed on 626 ASVs. For each instance of the model at internal node A , we set $\rho_A = 0.1$, $\alpha_\sigma = 0.5$, $\alpha_\eta = 1$, $\beta_\eta = 0.5$, and set β_σ to α_σ times a Taylor series approximation of $\hat{V}ar(F_j(A))$, based on the empirical proportions from the dataset of traversing down the left branch of node A .

4.2 Analysis of Patients Sampled 4 Times in 4 Weeks

To examine LoTGaP’s imputation, we took a subset of 21 allogeneic HCT patients with at least weekly measurements for 4 weeks after transplantation, and considered the 95 samples of these patients taken between 10 days prior to transplant up to 30 days after transplant (imputing all other points in between). For simplicity, we do not include any covariates in this example, and ran LoTGaP 2500 times, with 20% burn-in, thinning every other value.

Figure 4.1 plots the trajectories of 5 patients and 6 nodes that are part of *Lactobacillus* and *Enterococcus* paths (two genera recognized for their probiotic qualities [Fij14]). Table 4.1 describes the differences between the nodes in the context of *Lactobacillus* and *Enterococcus*, where “Always”, “Sometimes” and “Never” are unconditional on previous nodes (i.e. “Sometimes” *Lactobacillus* means not all paths

from the root to Lactobacillus leaves traverse through that node). In each subplot, the black line represents the median $F_j(A)$, the log-odds of traversal down the left branch of the node, with the dark gray and light gray representing the 50% and 95% confidence intervals, respectively. All empirical calculations of the log-odds are plotted as orange points (where points plotted below/above the minimum/maximum of the y -axis are points that empirically always traverse down the right/left branch of the node graphed, respectively). Discrepancies between the number of points for each patient occur when a node is reached at some time points but not others.

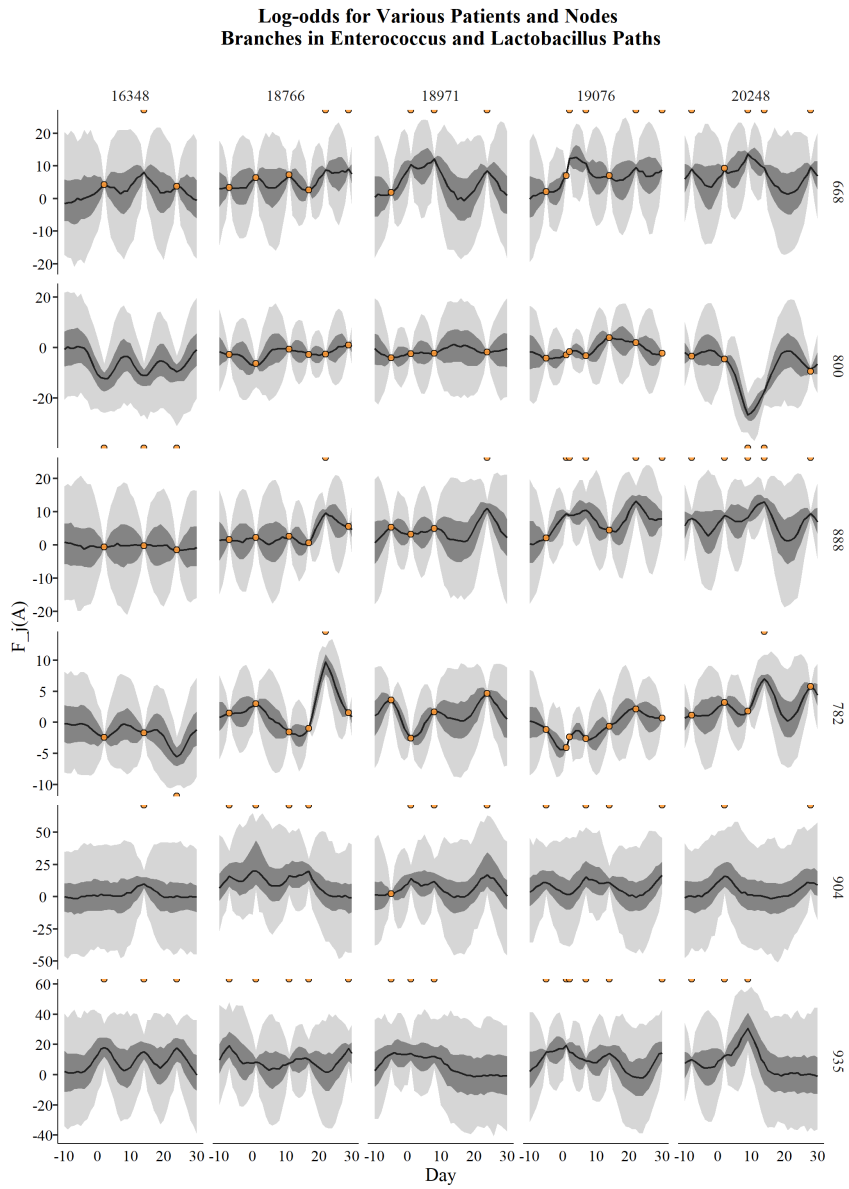


Figure 4.1: Trajectory of Lactobacillus and Enterococcus Internal Nodes

Node 668 is located towards the top of the tree. *Lactobacillus* and *Enterococcus* ASVs go through it, and both only move down the left path. We see that consistent with the data, subjects tend to have a slightly positive median trajectory, indicating above 50% probability of traversing down the *Lactobacillus* and *Enterococcus* portion of the tree, and larger periods of uncertainty come with (1) larger gaps between time points (2) large fluctuation in log-odds, or (3) extreme log-odds values (e.g., empirical log-odds of $-\infty$ or $+\infty$). However, information sharing narrows the confidence intervals even when all 3 of these factors are at play.

| Node Names | | | Node Properties | | | | | |
|------------|------------|-------------|-----------------|-------------|------------|-------------|-------------|--------------|
| Node | Left Child | Right Child | Lacto Node | Entero Node | Lacto Left | Entero Left | Lacto Right | Entero Right |
| 668 | 729 | 689 | Always | Always | Always | Always | Never | Never |
| 800 | 945 | 782 | Always | Sometimes | Never | Sometimes | Always | Sometimes |
| 888 | 855 | 901 | Always | Sometimes | Always | Sometimes | Never | Sometimes |
| 782 | 822 | 820 | Always | Sometimes | Never | Sometimes | Always | Never |
| 904 | 906 | 908 | Never | Sometimes | Never | Sometimes | Never | Sometimes |
| 935 | 923 | ASV 144 | Sometimes | Never | Sometimes | Never | Sometimes | Never |

Table 4.1: Node Split Descriptions in Figure 4.1

Nodes 800 and 888 are examples of nodes midway down the tree where *Lactobacillus* exclusively traverses down the right and left branches, respectively, while *Enterococcus* can traverse either. Once again, we see the largest uncertainty occur for patients with larger gaps and extreme log-odds values.

Node 782 is a “breaking point” node between *Lactobacillus* and *Enterococcus*, where *Lactobacillus* always traverses down the right branch, whereas *Enterococcus* always traverses down the left branch conditional on reaching 782. LoTGaP produces narrower confidence bands here, indicating less fluctuation at this breaking point. Intuitively, we would expect such a consequential node to be relatively stable, as a sudden imbalance towards one extreme of *Lactobacillus* and *Enterococcus* could be pathogenic.

Finally, *Enterococcus* and *Lactobacillus* touch only 904 and 935 respectively. In addition, these nodes do not partition either genus exclusively to one branch. Confidence bands are much larger here, to the point that 95% confidence intervals practically include the entire interval of $(0, 1)$ upon transforming the log-odds to a probability. This also gives insight into information sharing of LoTGaP, as we see much wider intervals for patient 16348 at nodes 904 and 935 than 800, despite exclusively having undefined odds ratios. When there are enough defined odds ratios across the pool of patients, the noise added to the Gaussian process cannot blow up.

Consequentially, this also implies that there could be value in modifying the priors of LoTGaP to allow for truncation of σ_A^2 , the technical variation in the Gibbs sampler

for node A . In the current version of LoTGaP, inference on nodes with mostly extreme log-odds diverges sometimes, especially when running it for thousands of iterations. We discuss this further in 4.4 and 5.

4.3 Analysis of Patients from Antiseptic Chlorhexidine Gluconate Study

To evaluate covariates, we take microbiome samples from a study evaluating the effectiveness of Chlorhexidine Gluconate (CHG) bathing ([GKR+21]). The sample consists of 105 patients, with 273 samples collected between 10 days prior to transplant up to 30 days after. There are 19 patients in study group 1 (no CHG bathing) with microbiome samples, 20 in study group 2 (< 50% of the time CHG bathing), 28 in study group 3 (50 – 75% CHG bathing), and 38 in study group 4 (75 – 100%).

We run LoTGaP 3750 times, with 20% burn-in and thinning down to every third iteration. We include the covariates of age, BMI, gender (with levels of Female, Male, and Unrecorded), race (with levels of Asian, Black, White, More than One, Unknown, and Unrecorded), transplant type (with levels of Allogeneic, Autologous, and Unrecorded), and study group (with levels 1-4).

8162 unique covariate-node combinations were fit as a part of the model. Figure 4.2 plots the 1000 posterior samples of the covariates (colored by iteration number and standardized to the log-odds scale) as well as 2.5th to 97.5th percentile interval (the black bands) for nodes 888 and 782, two of the six nodes graphed in 4.1.

Points in these plots are not clustered by color, suggesting that the mixing is appropriate with 20% burn-in and thinning to every third sample with these data. We can see that for both nodes, there are 2 variables whose coefficients have 95% credible intervals that do not cross 0 (Age, Study Group 4), and the sign of these coefficients is always negative within the 95% credible interval for node 888, and always positive for 782. We would expect the signs of the effects to differ between these two nodes since LoTGaP expresses covariates through mean shifts on the log-odds of traveling down the left branch of a node. As described in 4.1, all *lactobacillus* leaves go down the left branch at node 888, so being older and being in study group 4 (compared to the baseline of study group 1) are associated with a *lower* likelihood of traversing to *lactobacillus* leaves. On the other hand, all *lactobacillus* leaves traverse down the right branch at node 782, so being older and being in study group 4 (compared to the baseline of study group 1) are associated with a *higher* likelihood of traversing to *non-lactobacillus* leaves. We note that for node 888, the coefficient on Race Black has a 95% credible interval that does not include 0, and for node 782, the coefficient on BMI has a 95% credible interval that does not include 0 as well.

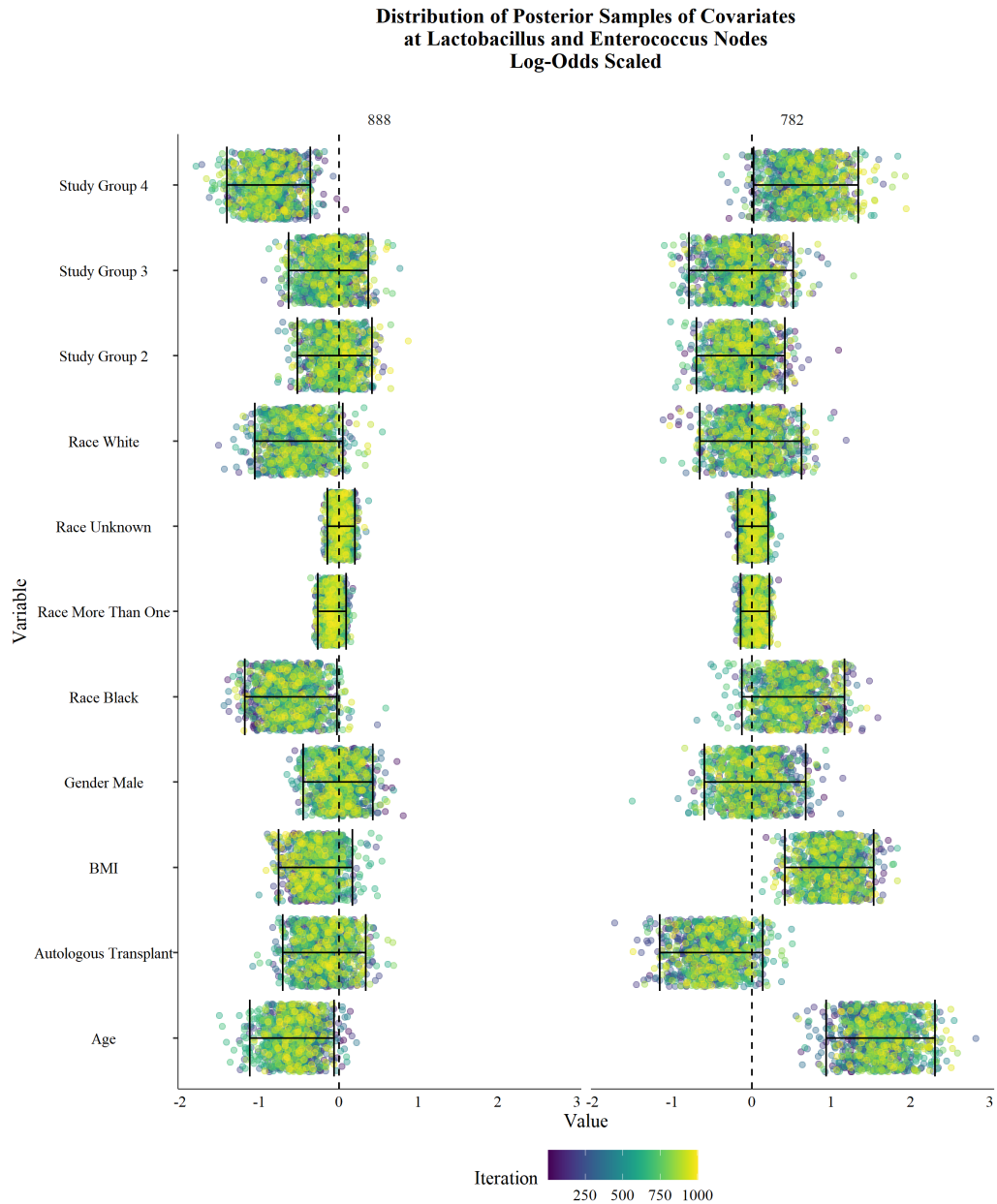


Figure 4.2: Distribution of Covariate Coefficients' Samples from LoTGaP run on CHG Study Samples

We also plot the 2.5th and 97.5th percentiles of all nodes by each variable in figure 4.3, highlighting nodes 888 and 782 in magenta and blue, respectively. Any points in the top left quadrant (colored in light orange) have 95% credible intervals that include 0 (the sign flips from the 2.5th percentile to the 97.5th percentile), whereas any points in the bottom left or top right quadrants (colored in light purple) have

95% credible intervals that are exclusively negative and positive, respectively. The majority of coefficient samples have 95% credible intervals containing 0, with Age, BMI, Study Group 4, and Male Gender being the variables that most frequently are associated with 95% credible intervals of coefficients that do not contain 0.

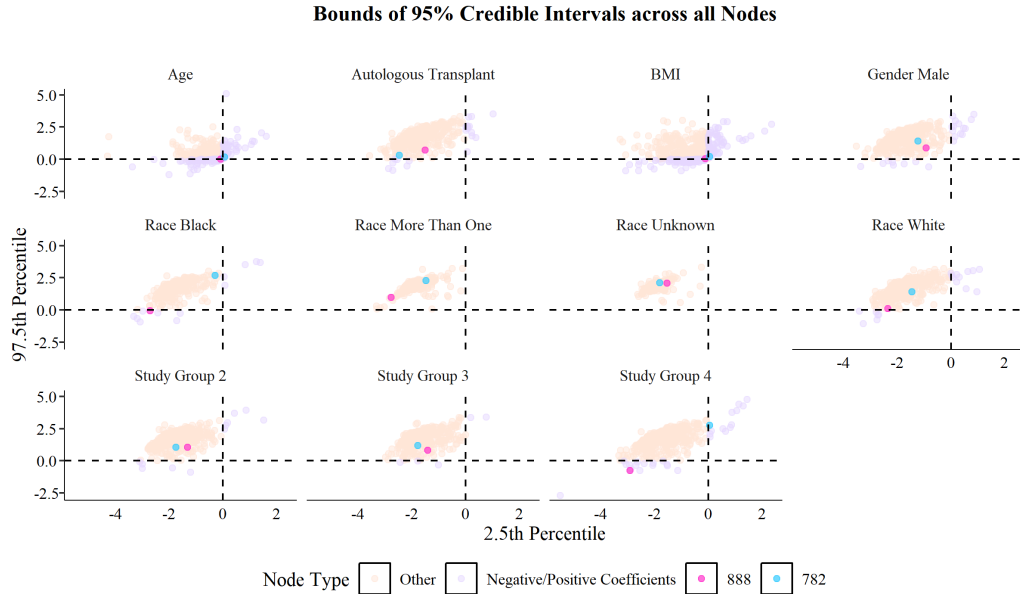


Figure 4.3: 2.5th vs. 97.5th Percentiles of Node-Covariate Samples

4.4 Run-Time Comparisons with TGP-CODA

As mentioned in 2.2.2, TGP-CODA is a different hierarchical time series model for the microbiome using Gaussian processes. However, TGP-CODA fits the model on a single patient at a time across all OTUs or ASVs, whereas LoTGaP fits the model on multiplate patients at a single internal node of a tree (but allows for parallelization across nodes). In addition, TGP-CODA uses Hamiltonian Monte Carlo methods implemented in STAN, whereas LoTGaP is a Gibbs sampler.

We ran both models on a local desktop, with no covariates or imputation included. LoTGaP was run 2500 times for each of the 625 internal nodes across the sample of 21 patients introduced in 4.2. TGP-CODA’s most recent implementation ([Äi19]) was run on patient 16348 across the 626 ASVs, one of the patients within the sample of 21 patients (and was graphed in 4.1).

LoTGaP took just under 38 minutes to complete its run across all nodes and 21 patients, whereas TGP-CODA took 2 hours and 54 minutes on a single patient. It is important to note though that LoTGaP exits early if (1) the model is unidentifiable due to the sample of subjects solely traversing down 1 branch at a node A , or (2)

the variance diverges to infinity within the sampler (due to very sparse counts of traversal down one branch of a node A across subjects), leading to parts of the model to become unidentifiable for A . Early exit status occurred on 242 of the 625 internal nodes, 148 being due to unidentifiability from the outset, and 94 occurring within the sampler itself. We discuss in section 5 potential remedies to address the divergence that is seen on the 94 nodes here. TGP-CODA also had divergence in the model 11.2% of the time, indicating that unidentifiability and divergence may be common issues across hierarchical Gaussian process models with these data.

In addition, TGP-CODA is not explicitly designed to model high-dimensional microbiome data, with authors running the model on 36 taxa in the original paper ([AMB18]). In future work, it would be a nice extension to generate synthetic data based on a tree with many nodes and a tree with fewer nodes and evaluate statistical power and run-time efficiency across both models for both trees. Nonetheless, the large difference in run-time performance underscores the potential value of utilizing the logistic tree in high-dimensional microbiome problems over other generalized logarithmic transforms.

Chapter 5

Discussion

Specialized modeling is required to analyze the trajectory of the human microbiome probabilistically with the advent of study designs of the microbiome containing inconsistently measured time points. We have derived an efficient hierarchical model that smooths out missing data across patients, that also allows for (1) highly granular microbiome data, and (2) covariate testing on the means of log-odds of branch traversal.

Within the current version of the model, there is room for modification. Developing an update step for ρ_A provides value in making the algorithm fully-Bayesian, though comes at the cost of the efficiency by not having all steps be Gibbs update steps, as well as forcing inversion of $K'(t, t' | \rho_A)$ at every iteration of the sampler. Therefore, we see this as a lower priority in areas of improvement.

However, it seems worthwhile to explicitly impose structure on the variance such that σ_A^2 does not blow up. A simple solution could be to bound σ_A^2 . A natural solution could be to truncate the inverse-Gamma prior on σ_A^2 , which still allows for efficient computation ([GSL92]). Additionally, while there is efficiency value brought by LoTGaP’s independent setup across nodes, changing the Gaussian kernel to be a Kronecker product between the existing time kernel and a separate covariance kernel between nodes could allow for improved inference, given that at the individual node level, identifiability can become an issue. Two potential avenues to explore are using a sparse covariance function, such as the Graphical Lasso ([FHT08]) to explicitly impose zero structure, or alternatively, design a covariance function based on distance within the phylogenetic tree to further inject biological information. Further sharing of information across nodes could be a valuable addition to LoTGaP, and careful consideration of covariance structure is required to define such a model.

In addition, for studies where data collection is set relative to an intervention, such as HCT patients at Duke Medical Center, it may be worthwhile to use a nonstationary kernel for the Gaussian process to explicitly partition temporal correlation pre-intervention and post-intervention. The Matérn covariance function is a nonstationary generalization of the exponential kernel used in LoTGaP ([PS06]), and can allow for further structure to LoTGaP induced by study design.

While this paper showed a simple example of posterior distributions of covariate coefficients within individual nodes across the tree over time, there is value towards developing hypothesis testing, specifically joint testing procedures across multiple nodes. There is an existing literature on tree hypothesis testing in genomics that has potential for extension here ([BFF⁺09]).

Finally, further development of ways to describe the model could help further

our understanding of LoTGaP and time-series modeling of the logistic tree on the microbiome relative to existing methods. Additional presentational tools that aggregate internal node probabilities to taxonomic ranks (such as families, genera, or species) and increased testing of differently sized trees (as well as different types of full binary trees related to the microbiome beyond phylogenetic trees) across different time series models are two key categories that can shape our comprehension of the microbiome. At this point, individual node identifiability is an issue in estimating uncertainty across entire paths to leaves of the tree, but with the other proposed changes for identifiability, this issue could start to disappear.

Appendix A

Pólya Gamma Augmentation

As mentioned in section 3.1.1, there are 2 key properties of Pólya-Gamma variables that we can leverage here:

$$p(\omega \mid b, \psi) = \frac{e^{-\omega\psi^2/2}p(\omega)}{\int_0^\infty e^{-\omega\psi^2/2}p(\omega)d\omega} \sim PG(b, \psi) \quad (\text{A.1})$$

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b}e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2}p(\omega)d\omega \quad (\text{A.2})$$

where $\kappa = a - b/2$, $p(\omega)$ is the probability density function of $PG(b, 0)$. Observe that A.2 is proportional to a binomial distribution. Setting $a = y_{jt}(A_\ell)$, $b = y_{jt}(A)$, $\psi = F_{jt}(A)$, $\omega = z_{jt}(A)$, we can alternatively express the joint distribution of $y_{jt}(A_\ell)$, $z_{jt}(A)$ conditional on $y_{jt}(A)$, $F_{jt}(A)$ as follows:

$$\begin{aligned} p(y_{jt}(A_\ell), z_{jt}(A) \mid y_{jt}(A), F_{jt}(A)) &= p(y_{jt}(A_\ell) \mid z_{jt}(A), y_{jt}(A), F_{jt}(A)) \\ &\quad \times p(z_{jt}(A) \mid y_{jt}(A), F_{jt}(A)) \\ &\propto p^{y_{jt}(A_\ell)}(1 - p)^{y_{jt}(A)} \frac{e^{-z_{jt}(A)(F_{jt}(A))^2/2}p(z_{jt}(A))}{\int_0^\infty e^{-z_{jt}(A)F_{jt}(A)^2/2}p(z_{jt}(A))dz_{jt}(A)} \end{aligned} \quad (\text{A.3})$$

$$\propto \frac{(e^{F_{jt}(A)})^{y_{jt}(A_\ell)}}{(1 + e^{F_{jt}(A)})^{y_{jt}(A)}} \frac{e^{-z_{jt}(A)(F_{jt}(A))^2/2}p(z_{jt}(A))}{\int_0^\infty e^{-z_{jt}(A)F_{jt}(A)^2/2}p(z_{jt}(A))dz_{jt}(A)} \quad (\text{A.4})$$

$$= 2^{-y_{jt}(A)} e^{\kappa_{jt}(A)F_{jt}(A)} e^{-z_{jt}(A)(F_{jt}(A))^2/2} p(z_{jt}(A)) \quad (\text{A.5})$$

$$= 2^{-y_{jt}(A)} e^{\kappa_{jt}(A)F_{jt}(A)} e^{-z_{jt}(A)(F_{jt}(A))^2/2} p(z_{jt}(A)) \quad (\text{A.6})$$

where $\kappa_{jt}(A) = y_{jt}(A_\ell) - y_{jt}(A)/2$. Observe that A.6 includes a normal distribution kernel for $F_{jt}(A)$, which is conjugate to 3.4. Thus, the augmentation allows for conjugate posterior updating of $F_{jt}(A)$ in a Gibbs sampler.

Appendix B

Full Conditionals

We use $-$ to denote all other covariates.

B.0.1 Full conditionals on mean parameters based on Multivariate Gaussian Conditionals

:

$$p(\mathbf{G}_j^T(A) | -) \sim \mathcal{N} \left(\Sigma_{\mathbf{G}_j} \left[\frac{1}{\sigma_A^2} I_{|\mathcal{T}_j|} (\mathbf{F}_j^T(A) - X_j \beta) \right], \Sigma_{\mathbf{G}_j} \right), \quad (\text{B.1})$$

$$\text{where } \Sigma_{\mathbf{G}_j} = \left(K_A(\mathcal{T}_j, \mathcal{T}_j | \eta_A, \rho_A)^{-1} + \frac{1}{\sigma_A^2} I_{|\mathcal{T}_j|} \right)^{-1}$$

$$p(\mathbf{F}_j^T(A) | -) \sim \mathcal{N} \left(\Sigma_{\mathbf{F}_j} \left[\frac{1}{\sigma_A^2} (\mathbf{G}_j^T(A) + X_j \beta) + \kappa_{jt}(A) \right], \Sigma_{\mathbf{F}_j} \right) \quad (\text{B.2})$$

$$\text{where } \Sigma_{\mathbf{F}_j} = \left(\frac{1}{\sigma_A^2} I_{|\mathcal{T}_j|} + z_j I_{|\mathcal{T}_j|} \right)^{-1}, \quad \kappa_{jt}(A) = y_{jt}(A_\ell) - y_{jt}(A)$$

$$p(\beta_A | -) \sim \mathcal{N} \left(\Sigma_{\beta_A} \left(\frac{1}{\sigma_A^2} \sum_{j=1}^J X_j^T (\mathbf{F}_j^T(A) - \mathbf{G}_j^T(A)) \right), \Sigma_{\beta_A} \right) \quad (\text{B.3})$$

$$\text{where } \Sigma_{\beta_A} = \left(\frac{1}{\tau^2} I + \frac{1}{\sigma_A^2} \sum_{j=1}^J X_j^T X_j \right)^{-1}$$

B.0.2 Full conditionals on variance terms based on Normal-Inverse Gamma Updates

:

$$p(\sigma_A^2 | -) \sim IG \left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\sigma, \frac{1}{2} \sum_{j=1}^J (\mathbf{F}_j(A) - \mathbf{G}_j(A) - X_j \beta_A)^T (\mathbf{F}_j(A) - \mathbf{G}_j(A) - X_j \beta_A) + b_\sigma \right) \quad (\text{B.4})$$

$$p(\eta_A^2 | -) \sim IG \left(\frac{\sum_{j=1}^J |\mathcal{T}_j|}{2} + \alpha_\eta, \frac{1}{2} \sum_{j=1}^J \mathbf{G}_j(A)^T K'_A(\mathcal{T}_j, \mathcal{T}_j | \eta_A^{(i)}, \rho_A)^{-1} \mathbf{G}_j(A) + b_\eta \right), \quad (\text{B.5})$$

$$\text{where } K'_A(\mathcal{T}_j, \mathcal{T}_j | \eta_A^{(i)}, \rho_A) = \frac{1}{\eta_A^2} K_A(\mathcal{T}_j, \mathcal{T}_j | \eta_A, \rho_A), \text{ so } k'(t, t' | \rho_A) = \exp(-\rho_A^{-2}(t - t')^2)$$

B.0.3 Full conditionals based on Pólya-Gamma augmentation (found in Appendix A)

:

$$p(z_{jt}(A) | -) \sim PG(y_{jt}(A), F_{jt}(A)) \quad (\text{B.6})$$

Bibliography

- [Ait82] J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, January 1982.
- [AMB18] Tarmo Äijö, Christian L Müller, and Richard Bonneau. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics*, 34(3):372–380, February 2018.
- [AMN⁺18] Jacob M. Allen, Lucy J. Mailing, Grace M. Niemi, Rachel Moore, Marc D. Cook, Bryan A. White, Hannah D. Holscher, and Jeffrey A. Woods. Exercise alters gut microbiota composition and function in lean and obese humans. *Medicine & Science in Sports & Exercise*, 50(4):747–757, 2018.
- [BCL⁺20] Francesco Biagini, Marco Calvigioni, Anna Lapomarda, Alessandra Vecchione, Chiara Magliaro, Carmelo De Maria, Francesca Montemurro, Francesco Celandroni, Diletta Mazzantini, Monica Mattioli-Belmonte, and et al. A novel 3D in vitro model of the human gut microbiota. *Scientific Reports*, 10(1), 2020.
- [BFF⁺09] Jorge R. Busch, Pablo A. Ferrari, Ana Georgina Flesia, Ricardo Fraiman, Sebastian P. Grynberg, and Florencia Leonardi. Testing statistical hypothesis on random trees and applications to the protein classification problem. *The Annals of Applied Statistics*, 3(2):542–563, June 2009.
- [Bis06] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.
- [BZB⁺04] Stéphanie Blanquet, Evelijn Zeijdner, Erick Beyssac, Jean-Philippe Meunier, Sylvain Denis, Robert Havenaar, and Monique Alric. A Dynamic Artificial Gastrointestinal System for Studying the Behavior of Orally Administered Drug Dosage Forms Under Various Physiological Conditions. *Pharmaceutical Research*, 21(4):585–591, April 2004.
- [CA19] Joana Costa and Arti Ahluwalia. Advances and Current Challenges in Intestinal in vitro Model Engineering: A Digest. *Frontiers in Bioengineering and Biotechnology*, 7:144, June 2019.
- [CL13] Jun Chen and Hongzhe Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1):418–442, March 2013.

- [CLC⁺11] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, 2011.
- [DMC⁺14] Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, and Peter J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, January 2014.
- [DMF⁺14] Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, 15(7):R89, 2014.
- [DPWF⁺10] Vicky De Preter, Karen Windey, Gwen Falony, Luc De Vuyst, and Kristin Verbeke. T2036 The Prebiotic Oligofructose-Enriched Inulin Affects the Faecal Metabolite Fingerprint: An In Vitro Analysis. *Gastroenterology*, 138(5), 2010.
- [FHT08] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.
- [Fij14] Sabina Fijan. Microorganisms with Claimed Probiotic Properties: An Overview of Recent Literature. *International Journal of Environmental Research and Public Health*, 11(5):4745–4767, May 2014.
- [GDRP⁺14] Julia K. Goodrich, Sara C. Di Rienzi, Angela C. Poole, Omry Koren, William A. Walters, J. Gregory Caporaso, Rob Knight, and Ruth E. Ley. Conducting a Microbiome Study. *Cell*, 158(2):250–262, July 2014.
- [GKR⁺21] Vinay K. Giri, Kristin G. Kegerreis, Yi Ren, Lauren M. Bohannon, Erica Lobaugh-Jin, Julia A. Messina, Anita Matthews, Yvonne M. Mowery, Elizabeth Sito, Martha Lassiter, Jennifer L. Saullo, Sin-Ho Jung, Li Ma, Morris Greenberg, Tessa M. Andermann, Marcel R.M. van den Brink, Jonathan U. Peled, Antonio L.C. Gomes, Taewoong Choi, Cristina J. Gasparetto, Mitchell E. Horwitz, Gwynn D. Long, Richard D. Lopez, David A. Rizzieri, Stefanie Sarantopoulos, Nelson J. Chao, Deborah H. Allen, and Anthony D. Sung. Chlorhexidine Gluconate Bathing Reduces the Incidence of Bloodstream Infections in Adults Undergoing Inpatient Hematopoietic Cell Transplantation.

Transplantation and Cellular Therapy, page S2666636721000026, January 2021.

- [GSL92] Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, 87(418):523–532, June 1992.
- [GWPGE16] Gregory B. Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It’s all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329, May 2016.
- [HES⁺20] Pernilla Lif Holgerson, Anders Esberg, Andreas Sjödin, Christina E. West, and Ingegerd Johansson. A longitudinal study of the development of the saliva microbiome in infants 2 days to 5 years compared to the microbiome in adolescents. *Scientific Reports*, 10(1), 2020.
- [HGZ18] Tao Hu, Paul Gallins, and Yi-Hui Zhou. A zero-inflated beta-binomial model for microbiome data analysis: ZIBB. *Stat*, 7(1):e185, 2018.
- [HHQ12] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS ONE*, 7(2):e30126, February 2012.
- [HLK10] Micah Hamady, Catherine Lozupone, and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal*, 4(1):17–27, January 2010.
- [HYJ⁺14] K. A. Harris, T. Yam, S. Jalili, O. M. Williams, K. Alshafi, T. Gouliouris, P. Munthali, U. Niriaian, and J. C. Hartley. Service evaluation to establish the sensitivity, specificity and additional value of broad-range 16S rDNA PCR for the diagnosis of infective endocarditis from resected endocardial material in patients from eight UK and Ireland hospitals. *European Journal of Clinical Microbiology & Infectious Diseases*, 33(11):2061–2066, 2014.
- [Ins] Broad Institute. DIABIMMUNE.
- [JSH⁺19] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):5029, December 2019.

- [KGS⁺15] Aleksandar D. Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöhö, Ismo Mattila, Harri Lähdesmäki, Eric A. Franzosa, Outi Vaarala, Marcus de Goffau, Hermie Harmsen, Jorma Ilonen, Suvi M. Virtanen, Clary B. Clish, Matej Orešič, Curtis Huttenhower, Mikael Knip, and Ramnik J. Xavier. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host & Microbe*, 17(2):260–273, February 2015.
- [KLW⁺12] Justin Kuczynski, Christian L. Lauber, William A. Walters, Laura Wegener Parfrey, José C. Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, January 2012.
- [Li15] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.
- [LSB⁺10] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, October 2010.
- [MAY18] Ahmed A. Metwally, Hani Aldirawi, and Jie Yang. A review on probabilistic models used in microbiome studies. *Communications in Information and Systems*, 18(3):173–191, 2018.
- [MMHV95] Mans Minekus, Phillipe Marteau, Robert Havenaar, and Jos H.J. Huis in’t Veld. A Multicompartmental Dynamic Computer-controlled Model Simulating the Stomach and Small Intestine. *Alternatives to Laboratory Animals*, 23(2):197–209, March 1995.
- [MSC⁺18] Deepanshi Mishra, Gita Satpathy, Rohan Chawla, Pradeep Venkatesh, Nishat Hussain Ahmed, and Subrat Kumar Panda. Utility of broad-range 16S rRNA PCR assay versus conventional methods for laboratory diagnosis of bacterial endophthalmitis in a tertiary care hospital. *British Journal of Ophthalmology*, 103(1):152–156, 2018.
- [MSQ⁺17] James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, and Rob Knight. Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2(1):mSystems.00162–16, e00162–16, February 2017.

- [Muk15] Sayan Mukherjee. Gaussian Process Regression. In *Probabilistic Machine Learning*, pages 57–59. American Mathematical Society, Durham, 2015. original-date: November 19, 2015.
- [MVWV93] K. Molly, M. Vande Woestyne, and W. Verstraete. Development of a 5-step multi-chamber reactor as a simulation of the human intestinal microbial ecosystem. *Applied Microbiology and Biotechnology*, 39(2):254–258, May 1993.
- [NCB18] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1):D8–D13, January 2018.
- [Nea97] Radford Neal. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, University of Toronto, January 1997.
- [PCK⁺18] Sarah C. Pearce, Heidi G. Coia, J. P. Karl, Ida G. Pantoja-Feliciano, Nicholas C. Zachos, and Kenneth Racicot. Intestinal in vitro and ex vivo Models to Study Host-Microbiome Interactions and Acute Stressors. *Frontiers in Physiology*, 9:1584, November 2018.
- [PGD⁺20] Jonathan U. Peled, Antonio L.c. Gomes, Sean M. Devlin, Eric R. Littmann, Ying Taur, Anthony D. Sung, Daniela Weber, Daigo Hashimoto, Ann E. Slingerland, John B. Slingerland, and et al. Microbiota as predictor of mortality in allogeneic hematopoietic-cell transplantation. *New England Journal of Medicine*, 382(9):822–834, 2020.
- [PRRL⁺14] Rachel Poretsky, Luis M. Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T. Konstantinidis. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE*, 9(4):e93827, April 2014.
- [PS06] Christopher J. Paciorek and Mark J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, August 2006.
- [PSW13] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, 108(504):1339–1349, December 2013.
- [Ras97] Carl Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*. PhD thesis, University of Toronto, Toronto, 1997.

- [RBF⁺20] Boyu Ren, Sergio Bacallado, Stefano Favaro, Tommi Vatanen, Curtis Huttenhower, and Lorenzo Trippa. Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. *The Annals of Applied Statistics*, 14(1):494–517, March 2020.
- [RW06] Carl Edward Rasmussen and Christopher K Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- [SDB⁺18] Justin D. Silverman, Heather K. Durand, Rachael J. Bloom, Sayan Mukherjee, and Lawrence A. David. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(1):202, December 2018.
- [SGW11] Patrick D. Schloss, Dirk Gevers, and Sarah L. Westcott. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE*, 6(12):e27310, December 2011.
- [SL17] Pixu Shi and Hongzhe Li. A model for paired-multinomial data and its application to analysis of data on a taxonomic tree. *Biometrics*, 73(4):1266–1278, December 2017.
- [SPS⁺12] Lakshmi Srinivasan, Jared M. Pisapia, Samir S. Shah, Casey H. Halpern, and Mary C. Harris. Can broad-range 16S ribosomal ribonucleic acid gene polymerase chain reactions improve the diagnosis of bacterial meningitis? a systematic review and meta-analysis. *Annals of Emergency Medicine*, 60(5), 2012.
- [SPT⁺20] Jonas Schluter, Jonathan U. Peled, Bradford P. Taylor, Kate A. Markey, Melody Smith, Ying Taur, Rene Niehus, Anna Staffas, Anqi Dai, Emily Fontana, Luigi A. Amoretti, Roberta J. Wright, Sejal Morjaria, Maly Fenelus, Melissa S. Pessin, Nelson J. Chao, Meagan Lew, Lauren Bohannon, Amy Bush, Anthony D. Sung, Tobias M. Hohl, Miguel-Angel Perales, Marcel R. M. van den Brink, and Joao B. Xavier. The gut microbiota is associated with immune cell dynamics in humans. *Nature*, 588(7837):303–307, December 2020.
- [SRMD20] Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020.
- [SSI⁺14] David Sims, Ian Sudbery, Nicholas E. Illott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, February 2014.

- [STNL⁺19] C. K. Stein-Thoeringer, K. B. Nichols, A. Lazrak, M. D. Docampo, A. E. Slingerland, J. B. Slingerland, A. G. Clurman, G. Armijo, A. L. C. Gomes, Y. Shono, A. Staffas, M. Burgos da Silva, S. M. Devlin, K. A. Markey, D. Bajic, R. Pinedo, A. Tsakmaklis, E. R. Littmann, A. Pastore, Y. Taur, S. Monette, M. E. Arcila, A. J. Pickard, M. Maloy, R. J. Wright, L. A. Amoretti, E. Fontana, D. Pham, M. A. Jamal, D. Weber, A. D. Sung, D. Hashimoto, C. Scheid, J. B. Xavier, J. A. Messina, K. Romero, M. Lew, A. Bush, L. Bohannon, K. Hayasaka, Y. Hasegawa, M. J. G. T. Vehreschild, J. R. Cross, D. M. Ponce, M. A. Perales, S. A. Giralt, R. R. Jenq, T. Teshima, E. Holler, N. J. Chao, E. G. Pamer, J. U. Peled, and M. R. M. van den Brink. Lactose drives *Enterococcus* expansion to promote graft-versus-host disease. *Science*, 366(6469):1143–1149, November 2019.
- [SWMD17] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, February 2017.
- [TBJ14] Olga Tanaseichuk, James Borneman, and Tao Jiang. Phylogeny-based classification of microbial communities. *Bioinformatics*, 30(4):449–456, February 2014.
- [TC19] Zheng-Zheng Tang and Guanhua Chen. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, October 2019.
- [TD20] Christine A. Tataru and Maude M. David. Correction: Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease. *PLOS Computational Biology*, 16(11):e1008423, November 2020.
- [VBE⁺15] Vaginal Microbiome Consortium (additional members), J Paul Brooks, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, Nihar U Sheth, Bernice Huang, Philippe Girerd, Jerome F Strauss, Kimberly K Jefferson, and Gregory A Buck. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15(1):66, December 2015.
- [VKd⁺16] Tommi Vatanen, Aleksandar D. Kostic, Eva d’Hennezel, Heli Siljander, Eric A. Franzosa, Moran Yassour, Raivo Kolde, Hera Vlamakis, Timothy D. Arthur, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Raivo Uibo, Sergei Mokurov, Natalya Dorshakova, Jorma Ilonen, Suvi M. Virtanen, Susanne J. Szabo, Jeffrey A. Porter, Harri Lähdesmäki, Curtis Huttenhower, Dirk Gevers, Thomas W.

- Cullen, Mikael Knip, and Ramnik J. Xavier. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, 165(4):842–853, May 2016.
- [WBPL91] W G Weisburg, S M Barns, D A Pelletier, and D J Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of Bacteriology*, 173(2):697–703, Jan 1991.
- [WWB90] David M. Ward, Roland Weller, and Mary M. Bateson. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65, May 1990.
- [XCJ⁺18] Jian Xiao, Li Chen, Stephen Johnson, Yue Yu, Xianyang Zhang, and Jun Chen. Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model. *Frontiers in Microbiology*, 9:1391, June 2018.
- [XPTX15] Lizhen Xu, Andrew D. Paterson, Williams Turpin, and Wei Xu. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLOS ONE*, 10(7):e0129606, July 2015.
- [Yon14] Ed Yong. Opinion | There Is No ‘Healthy’ Microbiome. *The New York Times*, November 2014.
- [YVS⁺16] Moran Yassour, Tommi Vatanen, Heli Siljander, Anu-Maaria Hämäläinen, Taina Hörkönen, Samppa J. Ryhänen, Eric A. Franzosa, Hera Vlamakis, Curtis Huttenhower, Dirk Gevers, Eric S. Lander, Mikael Knip, on behalf of the DIABIMMUNE Study Group, and Ramnik J. Xavier. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine*, 8(343):343ra81–343ra81, June 2016.
- [ZCS⁺18] Jun Zou, Benoit Chassaing, Vishal Singh, Michael Pellizzon, Matthew Ricci, Michael D. Fythe, Matam Vijay Kumar, and Andrew T. Gewirtz. Fiber-mediated nourishment of gut microbiota protects against diet-induced obesity by restoring il-22-mediated colonic health. *Cell Host & Microbe*, 23(1), Jan 2018.
- [Äi19] Tarmo Äijö. tare/GPMicrobiome, September 2019. original-date: 2016-04-22T13:53:41Z.