TIME-ALIGNED LATENT DIRICHLET ALLOCATION FOR LONGITUDINAL MICROBIOME DATA

TIME-ALIGNED LATENT DIRICHLET ALLOCATION FOR LONGITUDINAL MICROBIOME DATA

By

Sirikkathuge Fernando

A THESIS
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Master of Science (2023) McMaster University

(Statistics) Hamilton, Ontario, Canada

Title	:	Time-aligned Latent Dirichlet Allocation for Longitudinal Microbiome Data
Author	:	Sirikkathuge Fernando M.Sc in Statistics
Supervisor	:	Dr. Pratheepa Jeganathan
Number of pages	:	ix,63

Abstract

Microbial data exhibit dynamic characteristics driven by interactions among taxa and with experimental factors. This has led to an increased emphasis on longitudinal studies of microbial data due to their inter-dependencies. Emphasizing the temporal dynamics of microbial communities, rather than of individual taxa, provides valuable insights into the functionality of taxa. In the realm of identifying microbial communities, the probabilistic Latent Dirichlet Allocation (LDA) topic model has gained popularity (Sankaran & Holmes, 2019). This model is particularly applicable for analyzing multivariate, high-dimensional, and sparse data accommodating mixed membership in clusters.

This thesis introduces a time-aligned Latent Dirichlet Allocation (LDA), an extension of LDA for longitudinal microbiome data. Drawing inspiration from the work of Wang et al. (2021), our study aims to capture and analyze temporal changes in microbial communities.

The proposed time-aligned LDA method was implemented on gut microbial specimens obtained from pregnant women enrolled in the Be Healthy in Pregnancy (BHIP) study, both during pregnancy and at delivery. Subsequently, we conducted a comparative analysis with the traditional LDA approach using the hold-out specimen technique. Utilizing the time-aligned LDA alongside a mixed model, our findings indicates no discernible changes in microbial communities between treatment groups. Notably, the time-aligned LDA exhibited enhances sensitivity in identifying a greater number of microbial communities exhibiting significant

temporal dynamics compared to the standard LDA.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Pratheepa Jeganathan, for giving me the opportunity to conduct this research under her supervision. Her patience, guidance, constant suggestions, and encouragement helped me develop this research work. I truly respect and am indeed grateful for the wonderful supervision provided.

I would like to give my special thanks to Dr. Deborah Sloboda and Dr. Kate Kennedy from the Department of Biochemistry and Biomedical Sciences, McMaster University, for sharing the data set used in this thesis and for sharing necessary biological information for the conduct of this research.

I would also like to thank my examination committee members, Dr. Shui Feng and Dr. Ben Bolker, for their time and for providing comments for the improvement of the thesis.

Further, I would like to extend my thanks for the technical support given by SHARCNET and the Digital Research Alliance of Canada.

Finally, I am very grateful for the unconditional support I received from my loving husband, Tharaka, and my parents. Thank you for tolerating me at my worst and for always encouraging me to do my best. I could not have done this without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents an overview of longitudinal microbiome data. We identify the study objectives to be achieved and conclude with an outline work of the thesis.

1.1 Analysis of Microbiome Data

Microorganisms such as bacteria, viruses, protozoa, fungi generally co-exist in an environment (Micah et al., 2007, Symul et al., 2023). The development of modern technology of high-throughput sequencing (HTS) has enabled precise quantification of the microbial composition in a cost effective manner, which has resulted a massive improvement in microbiome-related research.

A microbiome study generally involves three main stages (Calle, 2019). First, microbial DNA are extracted from the collected specimens and sequenced using either amplicon sequencing or shot-gun metagenomics sequencing. The commonly used marker-gene for sequencing is 16S rRNA whereas in shot-gun metagenomic all genes present in the specimens are sequenced, allowing much greater resolution to identify microbial communities. In the second stage, the sequenced raw reads are processed and clustered into representative sequences referred as 'Operational

Taxonomic Units' (OTUs) (Edgar, 2013) or 'Amplicon Sequence Variants' (ASVs) (Callahan et al., 2016), commonly known as 'taxa'. We consider microbial communities as the taxa co-exist. This yields a count table containing the read counts or abundances of taxa for each specimen. Amplicon and metagenomic sequencing also result in a 'taxonomy table' which contains the taxonomic classifications of each taxon. Finally, the statistical and computational methods are used to investigate the distribution of microbial communities.

Many challenges have been identified in the microbial count table. The microbial count data are typically sparse, meaning they contain many zeros that can go up to approximately 90% (Joseph et al., 2013). This often leads to a skewed distribution for each taxon counts. Further, different subjects have different microbial compositions; even within a subject the microbial composition can be different in various human environments including vaginal, oral, gut. Therefore, there is high variability between and within subjects. The total count in the specimens is known as 'library size' which can also vary among specimens (McMurdie & Holmes, 2014). In addition, microbial count data have a large dynamic range varying from zero up to thousands and the variance of the data in different parts of the dynamic range are very different, resulting heteroskedasticity in data (Holmes & Huber, 2018).

Further, microbial count data parameters are compositional. Microbial data are converted to relative abundances, where the sum is equal to 1, meaning to compositional data, because of the unequal library sizes. Lastly, microbial data are multivariate as the number of taxa usually vary in thousands and there is dependence among taxa in the cooperative environment. As a result, existing standard statistical tools are limited with directly applying for analyzing amplicon and metagenomics sequencing data.

1.2 Longitudinal Microbiome Data

Longitudinal count data be $Y_t^{(s)}$ that denotes the non-negative integer response measured on subject s at time point t. Hence, longitudinal microbial count data be $\boldsymbol{Y_t^{(s)}}$ that denotes the integer response measured on subject s at time point t for V number of taxa. There is an increasing interest on studies focusing on the longitudinal microbial data because of the ability to gain more insights on within and between subject dynamics of microbial communities and their interactions with the experimental factors.

Longitudinal data poses more challenges in the statistical analysis. One should take into account the ordering of specimens within a subject. Moreover, characteristics of microbial data in Section 1.1 make the analysis more complicated in the context of a longitudinal study. For example, the sparsity becomes taxon and time specific (Kodikara et al., 2022). Also, the variability between subject counts may vary across different time points while varying for other factors. Further, the multivariate nature of $\boldsymbol{Y_t^{(s)}}$ might not be able to be compared over time as trends visible in relative abundance may not mirror the trends in actual abundance (Kodikara et al., 2022). In addition, the multivariate structure and auto correlation should be accounted for throughout all time points even though representation of dependence between taxa in a high-dimensional scenario is complex. Consequently, these challenges have led to the extension of statistical techniques to analyze longitudinal microbiome data.

1.3 Temporal Dynamics of Microbial Communities

One of the common goals in the analysis of longitudinal microbial data is modelling and identifying the temporal dynamics of microbial communities. How-

ever, taxon-wise longitudinal analysis is not suitable for this purpose since it often treats taxa as independent (Lugo-Martinez et al., 2019). Consequently, dynamic and network based methods were developed in the context of microbial count data; two such methods in popular are Generalized Lotka-Volterra (gLV) models (Stein et al., 2013, Fisher and Mehta, 2014, Bucci et al., 2016, Alshawaqfeh et al., 2017, Gibson and Gerber, 2018) and Dynamic Bayesian Network (DBN) models (McGeachie et al., 2014, Lugo-Martinez et al., 2019). The gLV models capture the changes in taxa composition while accounting for taxa interactions between themselves and with the external factors and external perturbations. The temporal dynamics are expressed in terms of ordinary differential equations that describe the ecological dynamics (gLV equations) which are later converted to a system of equations and solved. However, gLV models have high computational time because of the large number of parameters, which also lessens their utility for probabilistic inference (Ruiz-Perez et al., 2021). On the other hand, DBN models extends standard Bayesian networks to time series data which provides a probabilistic, non linear analogue to Auto-Regressive Integrated Moving-Average (ARIMA) and other linear models (Park et al., 2020). However, DBN models do not always reflect the actual taxa interactions and insufficient data may over-fit the model making predictions on non-existing interactions (Lugo-Martinez et al., 2019).

Dynamic models are designed to capture the temporal dynamics of longitudinal microbial count data while accounting for the interactions among taxa using a latent structure of taxa networks which remains unchanged over time. However, identifying temporal dynamics of microbial communities is important for understanding its metabolic and functional capabilities in human health (Symul et al., 2023).

In terms of identifying temporal dynamics of microbial communities, most of the studies have focused on taxon temporal patterns. For instance, Gerber

et al. (2012) proposed a multivariate time series clustering method called MC-TIMME (Microbiome Counts Trajectories Infinite Mixture Engine). MC-TIMME infers temporal latent patterns of longitudinal microbial count data using non-parametric Bayesian techniques with the Dirichlet process and assigns taxa to those inferred latent patterns. Moreover, the TIME web application introduced a new distance measure applicable to microbiome time series referred as 'TIME - dynamic time warping' (TIME-DTW) which uses a similarity measure to identify similar temporal patterns in taxa (Baksi et al., 2018). TIME calculates pair-wise TIME-DTW distances among taxa time series and uses it for hierarchical clustering. Coenen et al. (2020) also used agglomerative hierarchical clustering with complete linkage and partitioning around methoids (PAM) clustering for the microbiome time series analysis (Kaufman & Rousseeuw, 2009). Recently, Benincà et al. (2023) implemented Wavelet clustering to characterize community structure in human gut microbiome, which clusters time series based on similarities in their periodical patterns (Rouyer et al., 2008). Bodein et al. (2019) proposed a data-driven framework for clustering taxon temporal patterns. It involves fitting linear mixed model splines (LMMS) for temporal profiles of taxa. Then a dimension reduction on those splines is done using principal component analysis (PCA) and sparse principal component analysis (sPCA) and clustering is conducted based on PCA loadings.

1.4 Research Objectives

Many of the studies have investigated taxon temporal patterns. Instead, identifying microbial communities and how they change over time would provide more useful insights because there is dependence among taxa (Kodikara et al., 2022). Hence, the goal of this research is to study the temporal changes of microbial communities.

In identifying microbial communities, classical clustering or multivariate methods are no longer applicable because microbial data are high dimensional, multivariate count data characterized by heterogeneity and sparseness (Sankaran and Holmes, 2019, Jeganathan and Holmes, 2021). Further, the assumption that every microbial specimen belongs only to one community they made is often restrictive for human environments such as the gut (Holmes et al., 2012; Mao et al., 2020). Overcoming these problems, Sankaran and Holmes (2019) proposed the probabilistic topic model with latent Dirichlet allocation (LDA) in the context of microbial data. LDA was originally proposed in the context of document-text data by Blei et al. (2003). The main idea of LDA topic model is to consider every microbial specimen as a mixture of latent topics and every latent topic as a mixture of taxa, where every latent topic can be regarded as a microbial community. As a result, LDA uses the observed microbial count table to infer the latent topics present in microbial data by estimating the taxa distribution of topics and topic distribution of specimens using a Bayesian approach. Recently, Wang et al. (2021) introduced a longitudinal topic model which combines LDA and computational geometric representations to time evolving, complex and high-dimensional document-text data. We intend to implement a time-aligned topic model inspired by Wang et al. (2021) with modifications suitable to microbial data in order to achieve the goal of this study. To best of our knowledge, our study is the first application of a time-aligned topic model in the context of microbiome data.

Therefore, the primary goal of this research is to identify temporal changes in microbial communities by implementing a time-aligned topic model (Wang et al., 2021). In order to accomplish the research goal, we identify microbial communities using time-aligned LDA topic model. Then, we infer differential abundance across experimental factors using linear mixed models. Finally, we compare the results we obtain with the LDA model using hold-out sample technique.

1.5 Outline of the Thesis

This research focus on identifying the temporal changes of microbial communities. In order to accomplish the study goal, a time-aligned topic model is proposed. This dissertation consists of five main chapters, including the Introduction. Chapter 2 presents the statistical methods related to this study. Chapter 3 and 4 includes an example data set and exploratory data analysis and the application of the time-aligned LDA, respectively. Finally, Chapter 5 concludes the thesis by providing the limitations and suggestions for future research.

Chapter 2

Methods

This chapter provides a comprehensive statistical modelling facilitated during the conduct of this research. The main concepts discussed in this section are microbiome data components, Latent Dirichlet Allocation (LDA) model, time-aligned LDA model and linear mixed models.

2.1 Microbiome Data Structure

Microbial data are typically structured into three data tables. (1) The count matrix records specimens in rows, taxa in columns, and contains read counts in each cell (Table 1). (2) The sample data table is associated with the specimen information such as subject identity, sequencing batch, specimen type, and subject related experimental factors such as treatment group, demographics, having specimen identifiers in rows and specimen related variables on columns (Table 2). (3) The taxonomy table contains taxonomy levels such as species, kingdom, phylum, class, order, family, genus, having taxon identifiers in rows and taxonomy levels on columns (Table 3). One can also represent the hierarchical relationship between taxa in forms of a phylogenetic tree as demonstrated in Figure 1.

Table 1: Count matrix; matrix of V taxa in $D = \sum_{s=1}^{S} n_s$ specimens where n_s is the number of specimens collected from s^{th} subject over t time points for $s = 1, ..., S$. y_{dv} denote the abundance of taxon v in specimen d.

	Taxa_1	Taxa_2	...	Taxa_V
Specimen_1	y_{11}	y_{12}	...	y_{1V}
...	...	...	...	...
Specimen_{n_1}	$y_{n_1 1}$	$y_{n_1 2}$	...	$y_{n_1 V}$
Specimen_{n_1+1}	$y_{(n_1+1)1}$	$y_{(n_1+1)2}$	...	$y_{(n_1+1)N}$
...	...	...	...	...
Specimen_{n_2}	$y_{n_2 1}$	$y_{n_2 2}$	...	$y_{n_2 V}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Specimen_D	y_{D1}	y_{D2}	...	y_{DV}

Table 2: Sample data table.

	Subject ID	Time Point	Batch	Group
Specimen_1	Subject_1	1	1	Control
...	...	...	...	...
Specimen_{n_t}	Subject_1	t	1	Control
Specimen_{n_1+1}	Subject_2	1	2	Treatment
...	...	...	...	...
Specimen_{n_2}	Subject_2	t	2	Treatment
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Specimen_{n_S}	Subject_S	t	1	Control

Table 3: Taxonomy table.

	Kingdom	Phylum	Class	...	Genus
Taxa_1	Bacteria	Bacteroidetes	Bacteroidia	...	Parabacteroides
Taxa_2	Bacteria	Actinobacteria	Actinobacteria	...	Gardnerella
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Taxa_V	Bacteria	Firmicutes	Bacilli	...	FamilyXI

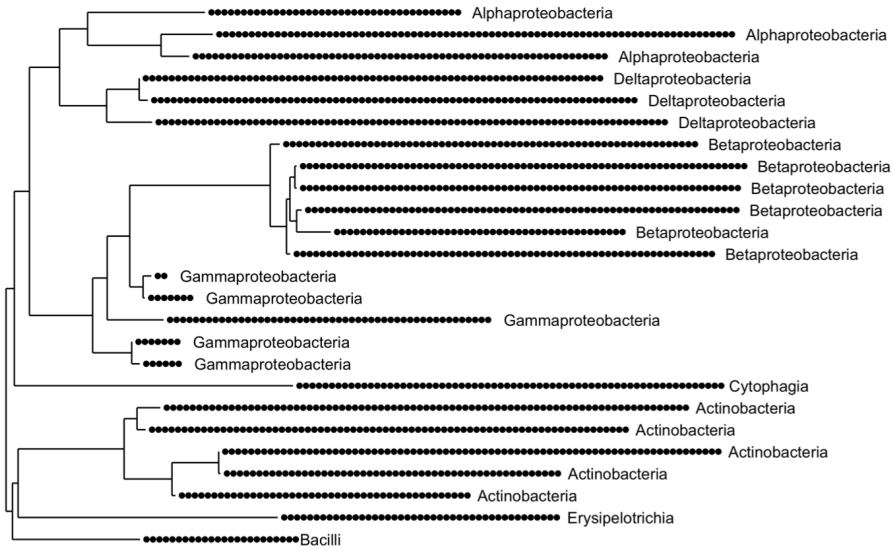


Figure 1: Phylogenetic tree; Black points at each node correspond to the number of specimens in which the corresponding taxa is present and tip labels are the class of the taxon. Source: Jeganathan and Holmes (2021).

In Table 2, let S be the number of subjects and $n_s; s = 1, .., S$ be the number of specimens collected from subjects over time points. Then the total number of specimens is,

$$D = \sum_{s=1}^{S} n_s. \tag{2.1}$$

Note that not all time points are available in all subjects. Let V be the number of taxa present in D specimens. Now, let $\mathbf{Y} = [y_{dv}] \in \mathbb{Z}^{D \times V}$ be the matrix of V taxa abundance in D specimens.

Note that,

- Every specimen d is associated with an V-variate vector of taxa abundance.

$$y_{d.} = (y_{d1}, y_{d2}, ..., y_{dV})^T. \tag{2.2}$$

- Further, the library size, the total taxa abundance in specimen d, is denoted by N_d.

$$N_d = \sum_{v=1}^{V} y_{dv}. \tag{2.3}$$

2.2 Statistical Transformations

Library sizes are highly variable among specimens, and are multiplicative factors which should be taken into account in statistical analysis (Jeganathan & Holmes, 2021). To remove the multiplicative effect of unequal library sizes, we calculate the library size scaling factor, denoted by N_d^* using the median-of-ratios algorithm (Anders & Huber, 2010). This method involves calculating the geometric mean of each taxon, dividing the abundance of each taxon by their corresponding geometric mean, and calculating the median of the ratios calculated for each specimen which is the library size scaling factor for that specimen.

We further consider variance stabilization for visualization purpose. Suppose we assume longitudinal microbiome count data follows a negative binomial distribution for each taxon, $y_{dv} \sim NB(\mu_v, k_v)$ where μ_v is the mean parameter and k_v is the dispersion parameter (Zhang et al., 2018). Then the Anscombe (1948) transformation for variance stabilization is given by, $y_{dv}^* = sin^{-1}(\sqrt{\frac{y_{dv}+c}{y_v-2c}})$. For $k_v > 2$ and large μ_v, $c = \frac{3}{8}$ and $\text{Var}(y_{dv}^*) = \frac{1}{4}\frac{1}{k_v-1/2}$.

2.3 Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic model introduced for discovering latent topics existing within a collection of documents (Blei et al., 2003). It is the foundation of all other variants of topic models. Here, we present the LDA topic model in the context of microbial count data (Sankaran & Holmes, 2019).

2.3.1 Generative Process of LDA

Blei et al. (2003) described LDA in text language, using "words" as the basic unit of discrete data, and "documents" and "corpora" which are the collection of words and collection of documents, respectively. The algorithm detailed below

uses words as taxa, documents as specimens and corpora as the microbial count matrix in Table 1 and a topic as a bacterial community (Sankaran & Holmes, 2019).

Let K be the pre-specified number of topics for the LDA.

The main idea in LDA is that it assumes every specimen is a mixture of latent topics, and every latent topic is a mixture of taxa as represented in Table 4 and 5.

Table 4: Topic mixture in D specimens; θ_{dk} denote the proportion of topic k in specimen d.

	Topic_1	Topic_2	...	Topic_K
Specimen_1	θ_{11}	θ_{12}	...	θ_{1K}
Specimen_2	θ_{21}	θ_{22}	...	θ_{2K}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Specimen_D	θ_{D1}	θ_{D2}	...	θ_{DK}

Table 5: Taxa mixture in topics; β_{kv} denote the proportion of taxon v in topic k.

	Taxa_1	Taxa_2	...	Taxa_V
Topic_1	β_{11}	β_{12}	...	β_{1V}
Topic_2	β_{21}	β_{22}	...	β_{2V}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Topic_K	β_{K1}	β_{K2}	...	β_{KV}

Let, θ_{dk} be the k^{th} topic proportion in d^{th} specimen, β_{kv} be the v^{th} taxa proportion in k^{th} topic. Then $\boldsymbol{\theta_d} = (\theta_{d1}, \theta_{d2}, ..., \theta_{dK})^T \in S^{K-1}$ and $\boldsymbol{\beta_k} = (\beta_{k1}, \beta_{k2}, ..., \beta_{kV})^T \in S^{V-1}$, where S^{K-1} and S^{V-1} are $(K-1)$ and $(V-1)$ simplex, respectively.

The Dirichlet distribution with hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_V)$ are used to generate the topic proportions in each specimen and the taxa proportions in each topic, respectively.

Let w_{dn} be the n^{th} sequence associated with one of the taxa in the d^{th} specimen and let z_{dn} be the topic assignment for the sequence w_{dn}.

For a given number of latent topics K, the generative process of LDA is defined as follows.

1. For each topic k,
   (a) Draw $\boldsymbol{\beta_k} \sim \text{Dirichlet}(\boldsymbol{\gamma})$.
2. For each specimen d,
   (a) Draw $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
   (b) For n^{th} sequence of specimen d,
      i. Draw $z_{dn}|\boldsymbol{\theta_d} \sim \text{Multinomial}(1, \boldsymbol{\theta_d})$,
      ii. Draw $w_{dn}|\boldsymbol{\beta_k}, z_{dn} \sim \text{Multinomial}(1, \beta_{z_{dn}})$.

We can summarize the above as follows,

$$\begin{aligned} w_{dn}|\boldsymbol{\beta_k}, z_{dn} &\overset{iid}{\sim} \text{Multinomial}(1, \beta_{z_{dn}}); d = 1, ..., D, ; n = 1, ..., N_d, \\ z_{dn}|\boldsymbol{\theta_d} &\overset{iid}{\sim} \text{Multinomial}(1, \boldsymbol{\theta_d}); d = 1, ..., D, ; n = 1, ..., N_d, \\ \boldsymbol{\theta_d} &\overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}); d = 1, ..., D, \\ \boldsymbol{\beta_k} &\overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\gamma}); k = 1, ..., K, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T$ is a K-dimensional vector with components $\alpha_i > 0$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_V)$ is a V-dimensional vector with components $\gamma_i > 0$.

We can create V taxon abundances in d^{th} specimen as $y_{dv} = \sum_{n=1}^{N_d} \mathbb{1}(w_{dn} \in v)$. Now by marginalizing over the topics the generative process of LDA can be written as in (2.5).

$$\begin{aligned} \boldsymbol{Y_{d.}}|(\boldsymbol{\beta_k})_{k=1}^K &\overset{iid}{\sim} \text{Multinomial}(N_d, \mathbf{B}\boldsymbol{\theta_d}); d = 1, ..., D, \\ \boldsymbol{\theta_d} &\overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}); d = 1, ..., D, \\ \boldsymbol{\beta_k} &\overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\gamma}); k = 1, ..., K, \end{aligned} \tag{2.5}$$

where $\mathbf{B} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_K})$ is $V \times K$ matrix having probability distribution of taxa in K topics.

2.3.2 Dirichlet Distribution

The Dirichlet distribution, denoted by $Dir(\boldsymbol{\alpha})$, is a multivariate generalization of the Beta distribution parameterized by a vector $\boldsymbol{\alpha} > 0$. The probability density function of a k-dimensional random vector $(x_1, ...x_K)^T; K \geq 2$ which has a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T$ is given as follows.

$$f(x_1, ...x_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{i=k}^{K} (\Gamma \alpha_k)} \prod_{k=1}^{K} x_k^{(\alpha_k - 1)}, \tag{2.6}$$

where $\sum_{k=1}^{K} x_k = 1$ and $x_k \in [0, 1]$, $\alpha_k > 0$ for $\forall k \in 1, ..., K$, and Γ is the gamma function, $x_k \in S^{(K-1)}$. $(x_1, ...x_K)$ belongs to a $(K-1)$ dimensional probability simplex that exists on a K-dimensional space.

The value of parameter $\boldsymbol{\alpha}$ determines where the density accumulates or spreads on the K-dimensional simplex. Figure 2 shows, for $0 < \boldsymbol{\alpha} < 1$ the density accumulates at the edges of the simplex, for $\boldsymbol{\alpha} = 1$ the density becomes more concentrated throughout the simplex uniformly and for $\boldsymbol{\alpha} > 1$ the density becomes more concentrated on the smaller subsets of the simplex. In addition, if $\alpha_k' s$ are same in $\boldsymbol{\alpha}$, then the density is symmetric. As a result, we set the Dirichlet hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ in LDA to a value less than 1 to generate mixtures different from each other in the topic proportion in specimens and taxa proportion in topics and avoid generating unrealistic topics (Jeganathan & Holmes, 2021).

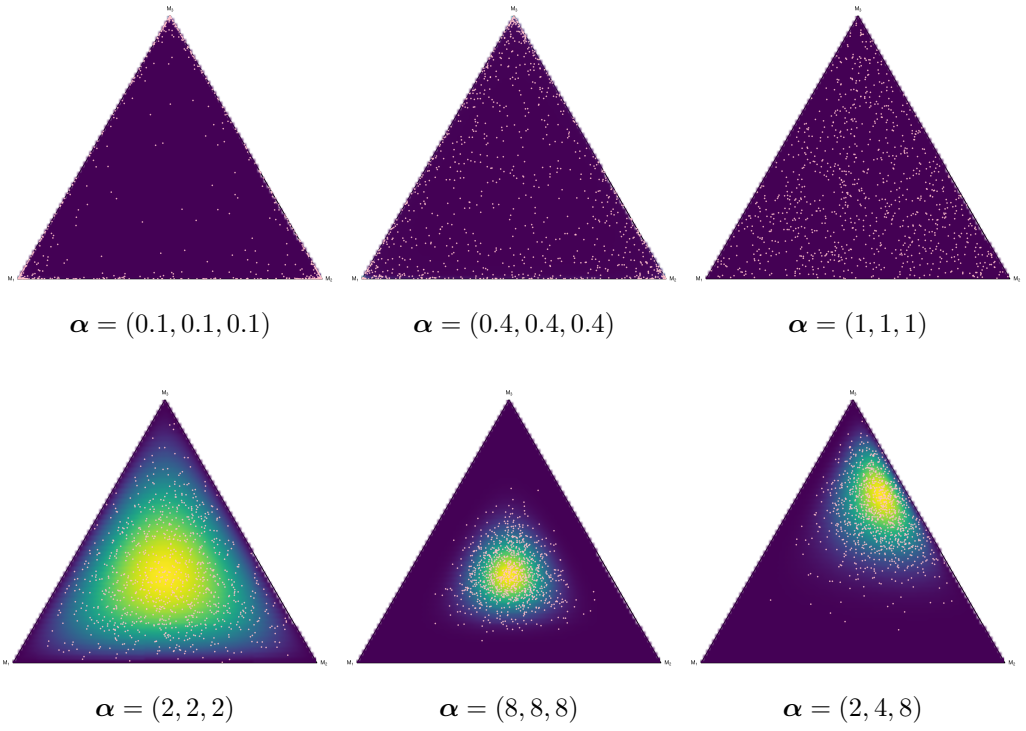


Figure 2: Distribution of 1000 points simulated from Dirichlet in a 3-dimensional space for different 3-dimensional $\boldsymbol{\alpha}$ parameter values.

2.3.3 Bayesian Inference

Given the number of latent topics K, the topic mixture in specimens $\boldsymbol{\theta_d}; d = 1, ..., D$ and the taxa mixture in topics $\mathbf{B} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_K})$ are the parameters that need to be estimated in the LDA.

The joint posterior distribution of $\boldsymbol{\theta_d}$ and $\mathbf{B}$ given $\boldsymbol{Y_{d.}}$ can be obtained using Bayes' theorem as follows.

$$\begin{aligned} P(\boldsymbol{\theta_d}, \mathbf{B}|\boldsymbol{Y_{d.}}, N_d, \boldsymbol{\alpha}, \gamma) &= \frac{P(\boldsymbol{Y_{d.}}|N_d, \boldsymbol{\theta_d}, \mathbf{B}) P(\boldsymbol{\theta_d}, \mathbf{B}|\boldsymbol{\alpha}, \gamma)}{P(\boldsymbol{Y_{d.}})} \\ &= \frac{P(\boldsymbol{Y_{d.}}|N_d, \boldsymbol{\theta_d}, \mathbf{B}) P(\boldsymbol{\theta_d}|\boldsymbol{\alpha}) P(\mathbf{B}|\gamma)}{P(\boldsymbol{Y_{d.}})} \\ &= \frac{P(\boldsymbol{Y_{d.}}|N_d, \boldsymbol{\theta_d}, \mathbf{B}) P(\boldsymbol{\theta_d}|\boldsymbol{\alpha}) P(\mathbf{B}|\gamma)}{\iint P(\boldsymbol{Y_{d.}}|N_d, \boldsymbol{\theta_d}, \mathbf{B}) P(\boldsymbol{\theta_d}|\boldsymbol{\alpha}) P(\mathbf{B}|\gamma)\, d\mathbf{B}\, d\boldsymbol{\theta}}. \end{aligned} \tag{2.7}$$

The second equation comes from assuming $\boldsymbol{\theta_d}$ and $\mathbf{B}$ are independent. The integral in the denominator in (2.7) is complicated as there are $D \times K$ number of θ

parameters and $K \times V$ number of β parameters involved. Hence, the joint posterior distribution of $\boldsymbol{\theta_d}$ and $\mathbf{B}$ is given by (2.8) below.

$$P(\boldsymbol{\theta_d}, \mathbf{B}|\boldsymbol{Y_{d.}}, N_d, \boldsymbol{\alpha}, \gamma) \propto P(\boldsymbol{Y_{d.}}|N_d, \boldsymbol{\theta_d}, \mathbf{B})P(\boldsymbol{\theta_d}|\boldsymbol{\alpha})P(\mathbf{B}|\gamma). \tag{2.8}$$

Next, the Markov Chain Monte Carlo (MCMC) algorithms can be used in order to draw samples from the posterior given by (2.8). We utilize Hamiltonian Monte Carlo (HMC) and its adaptive variant the No-U-Turn Sampler (NUTS) (Hoffman, Gelman, et al., 2014) for that purpose. For the sampler and computing we use Stan (Carpenter et al., 2017) which is a probabilistic programming language. It has an interface in R, RStan (Guo et al., 2020).

Hamiltonian Monte Carlo (HMC) and No-U-Turn Sampler (NUTS)

HMC is one of the MCMC algorithms that uses Hamiltonian dynamics to propose samples that follow a target distribution (Nishio & Arakawa, 2019).

Suppose, we are interest in drawing samples from the distribution $f(\theta)$ for a parameter θ. Typically, this is the posterior as shown in (2.9).

$$\begin{aligned} P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{P(X)}) \\ &\propto P(X|\theta)P(\theta) \\ &= f(\theta). \end{aligned} \tag{2.9}$$

HMC introduces auxiliary momentum variables ρ which is of the same dimension as θ with a multivariate normal density that does not depend on the parameter θ.

$$\rho \sim N(0, \Sigma). \tag{2.10}$$

The Hamiltonian, H is defined using the potential energy which depends on the parameter of interest, $V(\theta)$ and kinetic energy which depends on the momentum

parameter, $K(\rho)$ as follows.

$$H(\theta, \rho) = V(\theta) + K(\rho) \tag{2.11}$$

Then, the joint density function $f(\theta, \rho)$ has the form $f(\theta, \rho) \propto e^{-H(\theta,\rho)}$.

Next we can show,

$$\begin{aligned} H(\theta, \rho) &\propto -\log f(\theta, \rho) \\ &= -\log(f(\theta) f(\rho|\theta)) \\ &= -\log f(\theta) - \log f(\rho|\theta). \end{aligned} \tag{2.12}$$

As a result, potential energy is only determined by the target density, $V(\theta) = -\log f(\theta)$. Moreover, the kinetic energy is determined by the multivariate normal distribution; $K(\rho) = -\log f(\rho|\theta) = -\log f(\rho) = -\frac{1}{2}\rho^T \Sigma^{-1} \rho$. Therefore,

$$\begin{aligned} f(\theta, \rho) &\propto e^{(\log f(\theta) + \frac{1}{2}\rho^T \Sigma^{-1} \rho)} \\ &= f(\theta) e^{\frac{1}{2}\rho^T \Sigma^{-1} \rho}. \end{aligned} \tag{2.13}$$

Finally, we can obtain the target distribution by integrating $f(\theta, \rho)$ with respect to ρ. Now we show that

$$\begin{aligned} \int f(\theta, \rho) d\rho &= f(\theta) \int e^{\frac{1}{2}\rho^T \Sigma^{-1} \rho} d\rho \\ &= f(\theta). \end{aligned} \tag{2.14}$$

Thus, HMC generates samples from this joint distribution $f(\theta, \rho)$ and picks only θ in order to obtain the samples from the target distribution. Samples of (θ, ρ) are obtained with the use of Hamiltonian dynamics which describes the change of θ and ρ over time using the two differential equations given in (2.15) are called Hamilton's equations.

$$\begin{aligned} \frac{\mathrm{d}\theta}{\mathrm{d}t} &= \frac{\partial H}{\partial \rho} &= \frac{\partial K}{\partial \rho}, \\ \frac{\mathrm{d}\rho}{\mathrm{d}t} &= \frac{\partial H}{\partial \theta} &= \frac{\partial V}{\partial \theta}. \end{aligned} \tag{2.15}$$

To be specific, one should solve the Hamilton's equation and obtain the exact trajectory of state transition for θ and ρ from time t to time $t+s$, which allows to draw (θ, ρ) samples along those exact trajectories. However, in the practical usage, HMC approximates the trajectories in a discrete time setting, using the "Leapfrog" integrator which is a numerical integration algorithm. It involves applying integration steps of size ϵ over discrete time points $t = 1, .., L$ until t reaches L which is the number of integration steps in the leapfrog method. In the context of HMC, an integration step starts with a half-step update of ρ, followed by a full-step update of θ using updated ρ and ends again with a half-step update ρ using updated θ, as shown in (2.16).

$$\begin{aligned} \rho(t+\frac{\epsilon}{2}) &\leftarrow \rho(t) - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}, \\ \theta(t+\epsilon) &\leftarrow \theta(t) + \epsilon\rho(t+\frac{\epsilon}{2}), \\ \rho(t+\epsilon) &\leftarrow \rho(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}. \end{aligned} \tag{2.16}$$

Starting from (θ_0, ρ_0) and implementing L leapfrog steps of size ϵ, one can obtain the proposal state (θ^*, ρ^*). However, in order to account for numerical errors during integration resulting due to the time discretization, a Metropolis acceptance step is used, where either to accept (θ^*, ρ^*) with probability $min\{1, \exp(H(\theta_0, \rho_0) - H(\theta^*, \rho^*))\}$, or if (θ^*, ρ^*) is rejected, accept (θ_0, ρ_0).

No-U-Turn Sampler is a variant of HMC introduced to automate the tuning of these hyper-parameters because the performance of HMC is sensitive to the hyper-parameters L and ϵ (Hoffman, Gelman, et al., 2014).

We note that LDA does not account for the order of specimens in a subject. Thus,

LDA might be missing temporal changes in topic proportions and taxa proportions. Next, we propose to extend the LDA to account for temporal dynamics locally.

2.4 Time-aligned LDA

We propose a time-aligned topic model for longitudinal microbial data, inspired by Wang et al. (2021). The proposed model is a two stage approach, in which the first stage involves constructing new cohorts of specimens for each time point by subsampling and fitting independent topic models to each of the cohort, separately. The second stage involves aligning the topics in the cohorts.

2.4.1 Time-aligned LDA - Stage 1

Initially, a sub-sample of specimens, referred as a cohort, is constructed at each time point by conditional sampling of all specimens using a sampling distribution that is inversely proportional to the temporal adjacency of the specimens. This ensures that a cohort is a mixture of specimens of corresponding time point and nearby time points, providing a degree of temporal smoothing.

To elaborate the sub-sampling procedure, we suppose microbial count data corresponding to 5 specimens are collected over 4 time points.
Let, t_i indicate the i^{th} time point and C_i be the cohort created at i^{th} time point where $i = 1, 2, 3, 4$. For each cohort C_i, a 4-dimensional vector of exponential weights, denoted by $w_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})$ is calculated such that

$$w_{ij} = \begin{cases} 1 & \text{if } j = i, \\ 0.75^1 & \text{if } j = i-1 \text{ or } j = i+1, \\ 0.75^2 & \text{if } j = i-2 \text{ or } j = i+2, \\ 0.75^3 & \text{if } j = i-3 \text{ or } j = i+3. \end{cases} \tag{2.17}$$

Then, the sampling distribution is defined by normalizing the exponential weights w_{ij} for $i \neq j$ by their sum and is used to construct the sub-sample at time i. That is, when constructing the cohort at time point 1, C_1, exponential weights and the normalized weights denoted by w_1 and w_1^*, respectively are as in (2.18).

$$\begin{aligned} w_1 &= (w_{11} = 1, w_{12} = 0.75^1, w_{13} = 0.75^2, w_{14} = 0.75^3), \\ w_1^* &= (w_{11}^* = 1, w_{12}^* = 0.75^1/A_1, w_{13}^* = 0.75^2/A_1, w_{14}^* = 0.75^3/A_1) \\ &= (w_{11}^* = 1, w_{12}^* = 0.4324, w_{13}^* = 0.3243, w_{14}^* = 0.2432), \end{aligned} \tag{2.18}$$

where $A_1 = \sum_{j=1}^{4} w_{1j}$ for $i \neq j$.

Therefore, C_1 consists of 100% of specimens from time point 1, 43.24% of specimens from time point 2, 32.43% of specimens from time point 3, 24.32% of specimens from time point 4. This is similar to locally weighted regression (Cleveland & Devlin, 1988). When constructing the cohort at time point 2, C_2, exponential weights and the normalized weights denoted by w_2 and w_2^*, respectively are as in (2.19).

$$\begin{aligned} w_2 &= (w_{21} = 0.75^1, w_{22} = 1, w_{23} = 0.75^1, w_{24} = 0.75^2), \\ w_2^* &= (w_{21}^* = 0.75^1/A_2, w_{22}^* = 1, w_{23}^* = 0.75^1/A_2, w_{24}^* = 0.75^2/A_2) \\ &= (w_{21}^* = 0.3636, w_{22}^* = 1, w_{23}^* = 0.3636, w_{24}^* = 0.2727), \end{aligned} \tag{2.19}$$

where $A_2 = \sum_{j=1}^{4} w_{2j}$ for $i \neq j$.

We note that, C_2 consists of 36.36% of specimens from time point 1, 100% of specimens from time point 2, 36.36% of specimens from time point 3, 27.27% of specimens from time point 4 and so on (Figure 3). This sub-sampling procedure ensures a degree of smoothness over the cohorts generating at each time point in a manner that cohorts which are close in time likely to contain similar topics in specimens. Hence, this procedure is called temporal smoothing by sub-sampling

(Wang et al., 2021).

t_1	t_2	t_3	t_4
Specimen_{11}	Specimen_{12}	Specimen_{13}	Specimen_{14}
Specimen_{21}	Specimen_{22}	Specimen_{23}	Specimen_{24}
Specimen_{31}	Specimen_{32}	Specimen_{33}	Specimen_{34}
Specimen_{41}	Specimen_{42}	Specimen_{43}	Specimen_{44}
Specimen_{51}	Specimen_{52}	Specimen_{53}	Specimen_{54}

C_1

t_1	t_2	t_3	t_4
Specimen_{11}	Specimen_{12}	~~Specimen_{13}~~	~~Specimen_{14}~~
Specimen_{21}	~~Specimen_{22}~~	Specimen_{23}	~~Specimen_{24}~~
Specimen_{31}	Specimen_{32}	~~Specimen_{33}~~	~~Specimen_{34}~~
Specimen_{41}	~~Specimen_{42}~~	~~Specimen_{43}~~	~~Specimen_{44}~~
Specimen_{51}	Specimen_{52}	Specimen_{53}	Specimen_{54}

C_2

t_1	t_2	t_3	t_4
Specimen_{11}	Specimen_{12}	~~Specimen_{13}~~	~~Specimen_{14}~~
~~Specimen_{21}~~	Specimen_{22}	Specimen_{23}	~~Specimen_{24}~~
Specimen_{31}	Specimen_{32}	~~Specimen_{33}~~	Specimen_{34}
~~Specimen_{41}~~	Specimen_{42}	Specimen_{43}	~~Specimen_{44}~~
~~Specimen_{51}~~	Specimen_{52}	~~Specimen_{53}~~	~~Specimen_{54}~~

Figure 3: Demonstration of the sub-sampling procedure of 2 cohorts supposing 5 specimens collected over 4 time points.

After constructing the temporally smoothed cohorts of specimens for each time point, LDA topics models are implemented to each cohort independently, using the same prior distributions across all cohorts. These LDA models applied to time-localized smoothed cohorts are referred as local LDA models.

2.4.2 Time-aligned LDA - Stage 2

This stage involves combining the results obtained from local LDA models. In that regard, firstly, the topics of local LDA models are aligned. This is because, it is possible to observe topics in different order among the local LDA models since they are fitted independently.

In this study, we implement a local optimal alignment of topics using the cosine similarity. The cosine similarity between two vectors $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ is defined as

$$S_c(X,Y) = \frac{X.Y}{||X||||Y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{2.20}$$

Further, the topic distribution in specimens in the local LDA constructed for the first time point (cohort 1) is used as the reference for topic alignment. The topic alignment process is explained below.

1. To align topics in cohorts 1 and 2 (C_1,C_2),
   (a) Calculate the cosine similarity, S_c, between Topic 1 proportion in specimens in C_1 and each topic proportion in specimens in C_2.
   (b) Label the topic in C_2 that has the highest S_c as Topic 1.
   (c) Calculate S_c between Topic 2 in C_1 and each topic in C_2 except for the one labeled in (b).
   (d) Label the topic in C_2 that has the highest S_c as Topic 2.
   (e) Calculate S_c between Topic 3 in C_1 and each topic in C_2 except for the ones labeled in (b) and (d).
   (f) Label the topic in C_2 that has the highest S_c as Topic 3.
   (g) Repeat until all topics in C_2 are aligned.
2. To align topics in cohorts 2 and 3 (C_2,C_3),
   Treat aligned C_2 and C_3 as C_1 and C_2 in step (1), respectively and repeat steps in (1).
3. To align topics in cohorts 3 and 4 (C_3,C_4),
   Treat aligned C_3 and C_4 as C_1 and C_2 in step (1), respectively and repeat steps in (1).

After the topic alignment, a time-aligned topic distribution in specimens is obtained in a way that, for a specimen corresponding to the i^{th} time point, the topic distribution in specimens obtained from the local LDA implemented on i^{th} cohort C_i, is the topic distribution for that specimen.

2.5 Mixed Models

2.5.1 Negative Binomial Mixed Model for Taxa Abundance

With the purpose of identifying taxa with differential abundance among the experimental factors including treatment groups, Negative Binomial Mixed Models (NBMM) were fitted for each taxon separately.

We consider D number of specimens and V number of taxa. Denote the abundance of any taxa by Y_v which is assumed to follow a negative binomial distribution (Zhang et al., 2018). The probability density of Y_v is as follows.

$$f(y_v) = \frac{\Gamma(y_v + \phi_v)}{\Gamma(\phi_v) y_v!} . \left(\frac{\phi_v}{\mu_v + \phi_v}\right)^{\phi_v} . \left(\frac{\mu_v}{\mu_v + \phi_v}\right)^{y_v} ; y_v = 0, 1, ..., \mu_v \geq 0, \phi_v \geq 0. \tag{2.21}$$

where μ_v and ϕ_v are the mean and dispersion parameters, respectively and $\Gamma(p) = \int_0^\infty t^{(p-1)} e^{-t} dt$ is the gamma function.

Thus, $E(Y_v) = \mu_v$ and $Var(y_v) = \sigma_v^2 = \mu_v + \frac{\mu_v^2}{\phi_v}$.

The NBMM for taxa abundance Y_v with p number of fixed-effects (including the intercept) and q number of random-effects where the vector of random effects is denoted by $\mathcal{B}$ is given by,

$$(\boldsymbol{Y_v} | \mathcal{B} = \boldsymbol{b}) \sim \text{NB}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b}, \sigma^2 \mathbf{W}^{-1}), \tag{2.22}$$

where $\boldsymbol{Y_v}$ is the D-dimensional response vector of taxa abundance, $\mathbf{X}$ is the $D \times p$ design matrix of fixed-effects, $\boldsymbol{\beta}$ is the corresponding coefficient vector, $\mathbf{Z}$ is the $D \times q$ design matrix of random-effects, $\boldsymbol{b}$ is the corresponding coefficient vector, $\mathbf{W}$ is a diagonal matrix of known prior weights and σ is the scale parameter. Here is the description of $\boldsymbol{Y_v}$, $\mathbf{X}$, $\mathbf{Z}$, $\boldsymbol{\beta}$ and $\boldsymbol{b}$.

$$\boldsymbol{Y_v} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{pmatrix}, \mathbf{X}= \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{D1} & ... & x_{Dp} \end{pmatrix}, \mathbf{Z}= \begin{pmatrix} z_{11} & ... & z_{1q} \\ z_{21} & ... & z_{2q} \\ \vdots & \ddots & \vdots \\ z_{D1} & ... & z_{Dq} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & ... & \beta_p \end{pmatrix}^T, \boldsymbol{b} = \begin{pmatrix} b_1 & b_2 & ... & b_q \end{pmatrix}^T.$$

In general, the vector of the random effects is assumed to follow a multivariate normal distribution, $\boldsymbol{b} \sim N(0, \psi)$ where ψ is a positive semi-definite variance-covariance matrix that determines the dependence between random effects. In this study, for simplicity, the case where the random effects are independent is assumed, i.e.: $\boldsymbol{b} \sim N(0, \mathbf{I}_q)$. Further, we assume unit weights for $\mathbf{W}$, that is $\mathbf{W} = \mathbf{I}_q$.

2.5.2 Linear Mixed Model for Topic Proportion

In order to identify latent topics derived from LDA that are differentially abundant between the experimental factors, Linear Mixed Models (LMM) were fitted for each topic separately.

Recall that we consider D number of specimens, V number of taxa and K number of latent topics. Denote the topic proportion of any topic by P_k. We fit a LMM for the logarithm of topic proportions.

The LMM for the logarithm of topic proportion P_k for each topic k with p number of fixed-effects (including the intercept) and q number of random-effects where the vector of random effects is denoted by $\mathcal{B}$ is given by,

$$(\log(\boldsymbol{P_k})|\mathcal{B} = \boldsymbol{b}) \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b}, \sigma^2\mathbf{W}^{-1}) \tag{2.23}$$

where $\boldsymbol{P_k}$ is the D-dimensional response vector of k^{th} topic proportion, and $\mathbf{X}$, $\mathbf{Z}$,

$\boldsymbol{\beta}$, $\boldsymbol{b}$, $\mathbf{W}$ and σ are same as given in (2.22).

As similar to Section 2.4.1, the independent random effects and unit weights are assumed.

In the mixed models for taxa abundance and topic proportions mentioned above, models were fitted using the `glmer.nb` and `lmer` functions in `lme4` package (Bates et al., 2009) and the maximum likelihood estimation with Laplace approximation and Nelder Mead optimizer was used.

Chapter 3

Example Dataset and Application

This purpose of this chapter is to offer an insight on an example longitudinal microbiome data set that is considered in the application of the proposed time-aligned LDA. Hence, Chapter 3 uncovers details regarding the data set including the process of data preparation and data exploration.

3.1 Example Data set

The microbial data used are from the 'Be Healthy in Pregnancy'(BHIP) which is a randomized two-arm study, (Atkinson et al., 2022). It was designed to test the efficacy of a novel treatment which provided structured and monitored nutrition (high dairy protein diet and individualized energy intake), exercise (walking of 10,000 steps/day) with bi-weekly counselling compared to the regular care as per Health Canada recommendations, through out the pregnancy. Data consists of gut microbial specimens of 63 pregnant women in southern Ontario, Canada, collected at each trimester and one time after the delivery, resulting in 191 total specimens and gut microbial specimens of 50 infants collected one time within 6 months after the delivery, resulting 50 total specimens. The sampling schedule is demonstrated

in Figure 4 and Table 6 provides a summary of the variables in sample data.

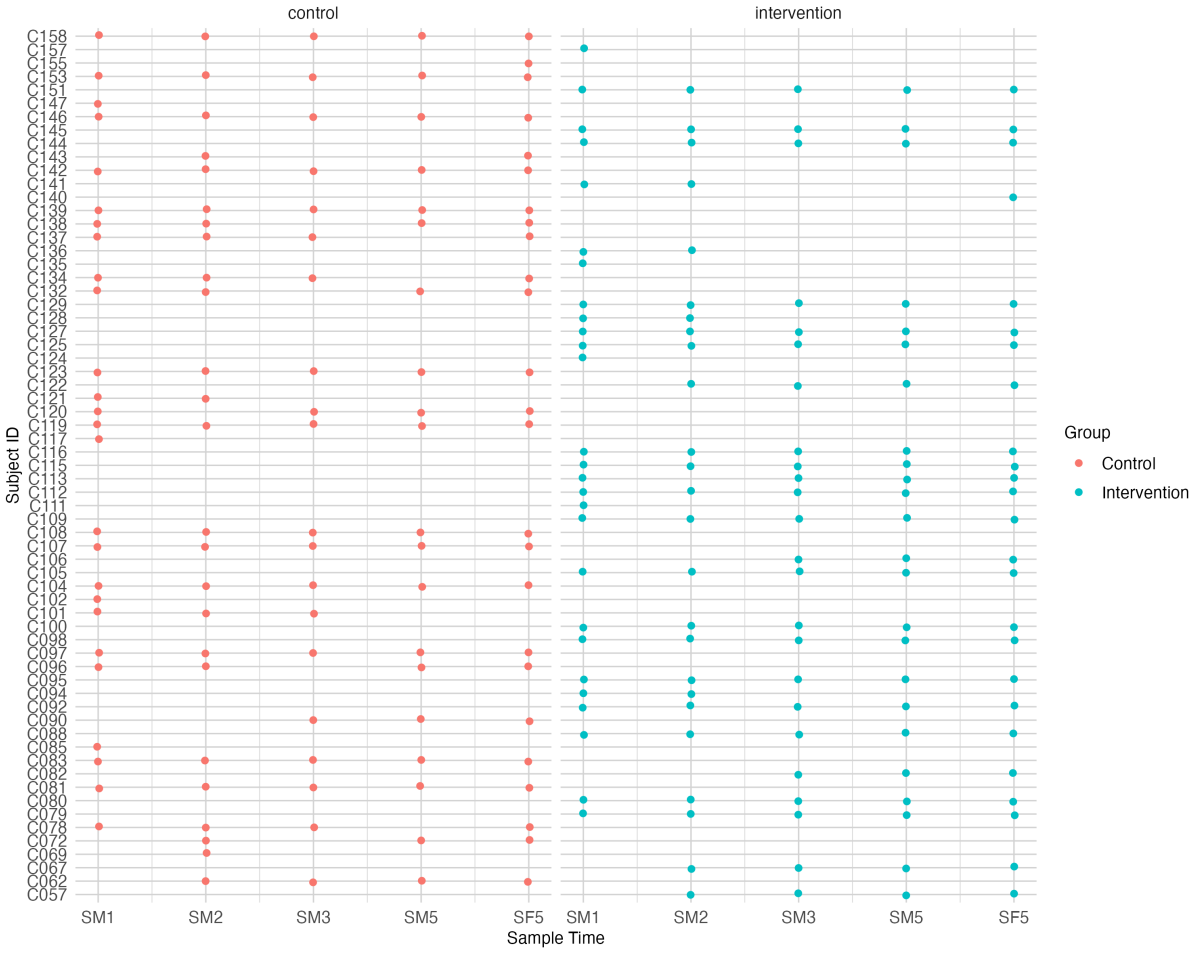


Figure 4: Sampling schedule. Participants were grouped (colour) into one of control group or intervention group. For maternal specimens, SM1 = first trimester; SM2 = second trimester; SM3 = third trimester; SM5 = 6 months postpartum and for infant specimens, SF5 = infant at 6 months of age.

Table 6: Description of sample variables.

Variable Name	**Description**
participantID	Unique identity for the participant
sampleID	Unique identity for the specimen
sampleTime	Time at point the specimen were collected. (Maternal : SM1 = first trimester; SM2 = second trimester; SM3 = third trimester; SM5 = 6 months post-partum; Infant : SF5 = infant at 6 months of age)
illuminaRun	Sequencing run
group	Treatment group (intervention/control)
BMI_SM0	Pre-pregnancy BMI (kg/m^2)
gravidity	Number of previous pregnancies
dG	Days of gestation (length of pregnancy in days)
weight_gain	Total gestational weight gain (in kgs)
birthMode	Delivery mode (Spontaneous/Forceps/Vacuum extraction/Emergency caesarean section/Scheduled caesarean section)
modeSimple	Simplified birth mode (vaginal = Spontaneous/Forceps/Vacuum extraction; cs = Emergency caesarean section/Scheduled caesarean section)
sex	Infant sex (female/male)
gwg_kg_wk	Gestational weight gain rate (kg/week)
BMI_cat_0	Pre-pregnancy BMI category (underweight = BMI<18.5; optimal = BMI within (18.5–24.9); overweight = BMI within (25.0–29.9); obese = BMI> 29.9)
gain_cat	Gestational weight gain category (below/ wihtin/ above the range by pBMI according to Institute of Medicine guidelines)
grava	Gravidity category (0 = primigravid; 1 = multigravid)
EADP	Excessive adiposity during pregnancy (Yes = pBMI >25 or above gain_cat; No = otherwise)

3.2 Exploratory Data Analysis

Because maternal and infant microbial communities are hugely different, we consider maternal gut microbial specimens from 63 pregnant women in the thesis. Participant were equally distributed among treatment groups (control n=31; intervention n=32). With respect to the levels of gravidity about 50% were categorized as multigravid. The number of participants falling in to the levels of pre-pregnancy body mass index (pBMI) is approximately similar between the treatment groups where about 60-70% were categorized as optimal, about 20-30% as overweight and about 10% as obese (Table 7).

Table 7: Experimental factors.

Demographic Characteristic	**Control (n=31)**	**Intervention (n=32)**
	Mean(SD)	**Mean(SD)**
Pre-pregnancy BMI (kg/m^2)	23.54 (4.05)	23.87 (3.90)
Gravidity	1.25 (1.57)	0.85 (1.07)
	N(%) (Mean,SD)	**N(%) (Mean,SD)**
Pre-pregnancy BMI		
Underweight	0 (0) (0.00,0.00)	1 (3) (17.37,0.00)
Optimal	19 (61) (22.25,1.79)	22 (69) (22.29,1.23)
Overweight	9 (29) (26.72,1.14)	6 (19) (26.90,1.37)
Obese	3 (10) (32.38,1.28)	3 (9) (35.45,3.48)
Gravidity		
0	15 (48)	17 (53)
1	15 (48)	14 (44)
Missing	1	1

The observed count table contains information on 4019 taxa in 191 specimens. We filtered taxa by removing non-bacterial taxa, the ones belonged to kingdom *Eu-*

karyota, family *Mitochondria*, order *Chloroplast*, or no assigned phylum (n=235). We also removed low prevalence taxa which are present in less than 5% of samples (n=2915). Further, there were 10 specimens which corresponds to a subject where that subject has only that specimen. Therefore, there is no longitudinal data available for that subject. Since, this study focuses on temporal dynamics, we filtered those 10 specimens. The reduced count table contains information on 869 taxa in 181 specimens.

Initially, we explored within microbial community diversity in participants throughout pregnancy and post-delivery using Shannon diversity (SD) index (Shannon, 1948). It is an alpha diversity metric which estimates the diversity of microorganisms within a community. We noticed SD index increased over time on the intervention group subjects which is actually caused by one outlier appearing in the first time point in that group (Figure 5). However, Figure 6 demonstrates an increasing trend in the SD index for the subjects with optimal pBMI and a decreasing trend of the same for subjects with over-weight pBMI over time in the intervention group. Also, a similar increasing and decreasing behaviour is observed

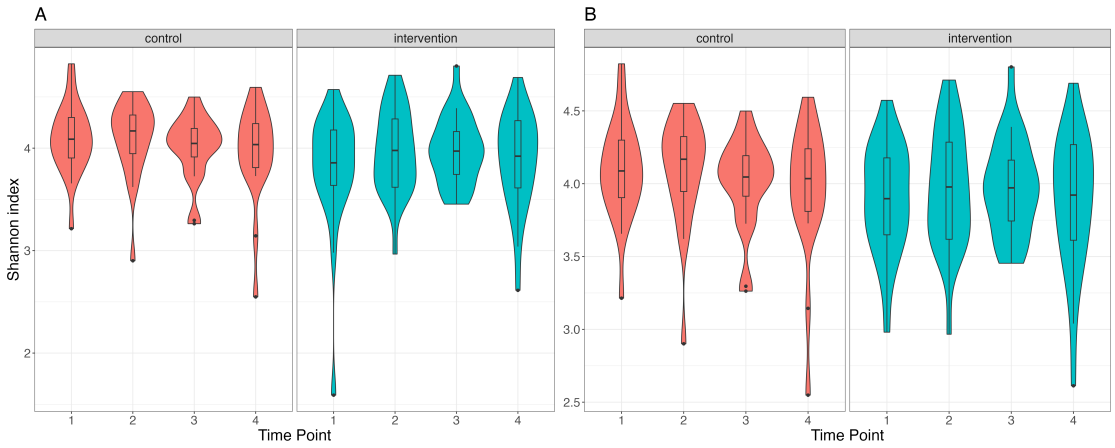


Figure 5: Distribution of Shannon index throughout the four-time points between treatment groups; A: Before removing the outlier, B: After removing the outlier.

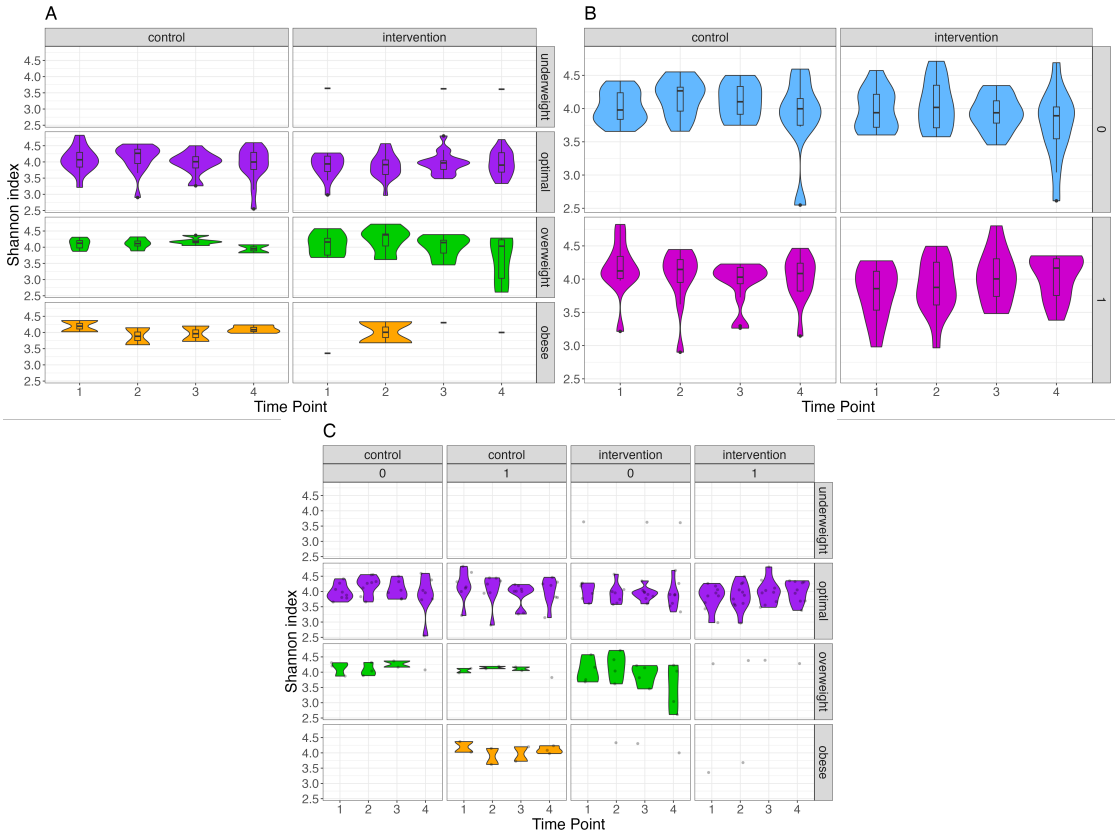


Figure 6: Distribution of Shannon index throughout the four-time points between treatment groups and; A: pBMI, B: Gravidity C: Both pBMI and gravidity.

in SD index but with respect to with and without previous pregnancies (Figure 6:(B)). More interestingly, looking into further we can see that, in the intervention group, the decreasing trend for the subjects with no previous pregnancies is actually corresponding to over-weight subjects and the increasing trend for subjects with previous pregnancies corresponds to optimal weight subjects (Figure 6:(C)).

Next, ordination methods were applied based on the Bray-Curtis and Jaccard distances in order to identify outliers, clusters and gradients of specimens of microbial distribution in low dimensions.Figures 7 and 8 shows there is no outliers and clusters, but there are changes in the specimen ordination over time but not among treatment groups.

The alpha diversity and beta diversity based ordination limited in uncovering temporal changes in taxa.

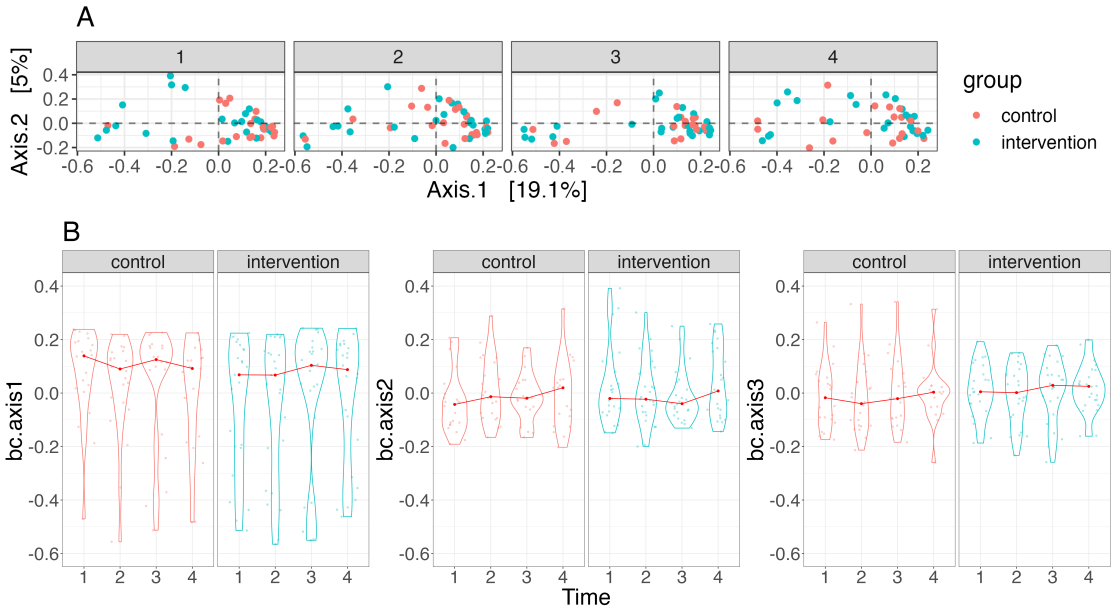


Figure 7: Ordination using Bray-Curtis distances; A: Ordination plot over time, B: Ordination axes values over time.

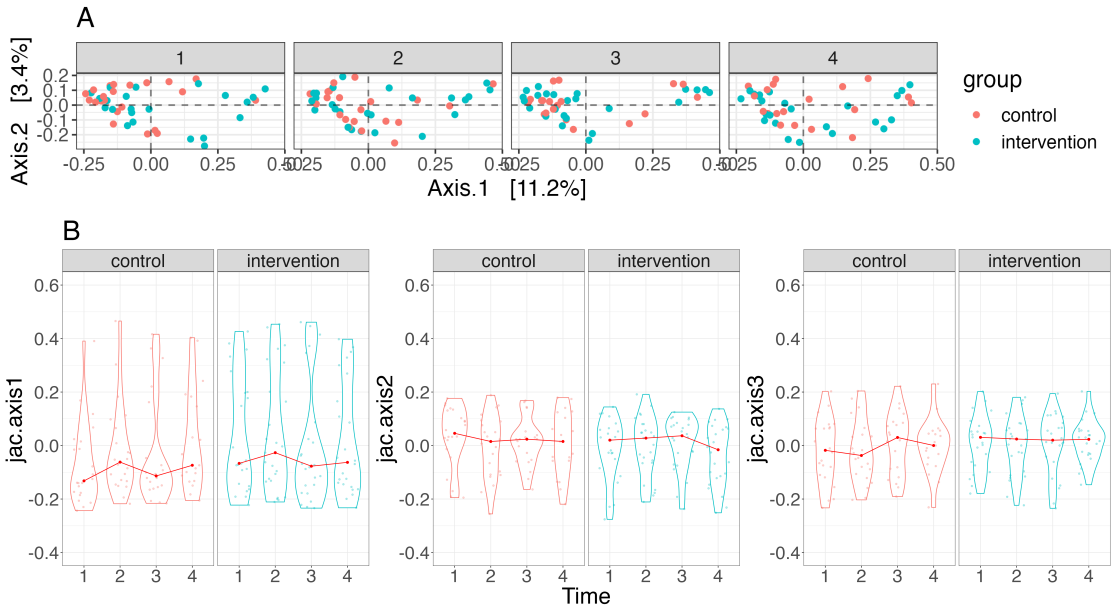


Figure 8: Ordination using Jaccard distances; A: Ordination plot over time, B: Ordination axes values over time.

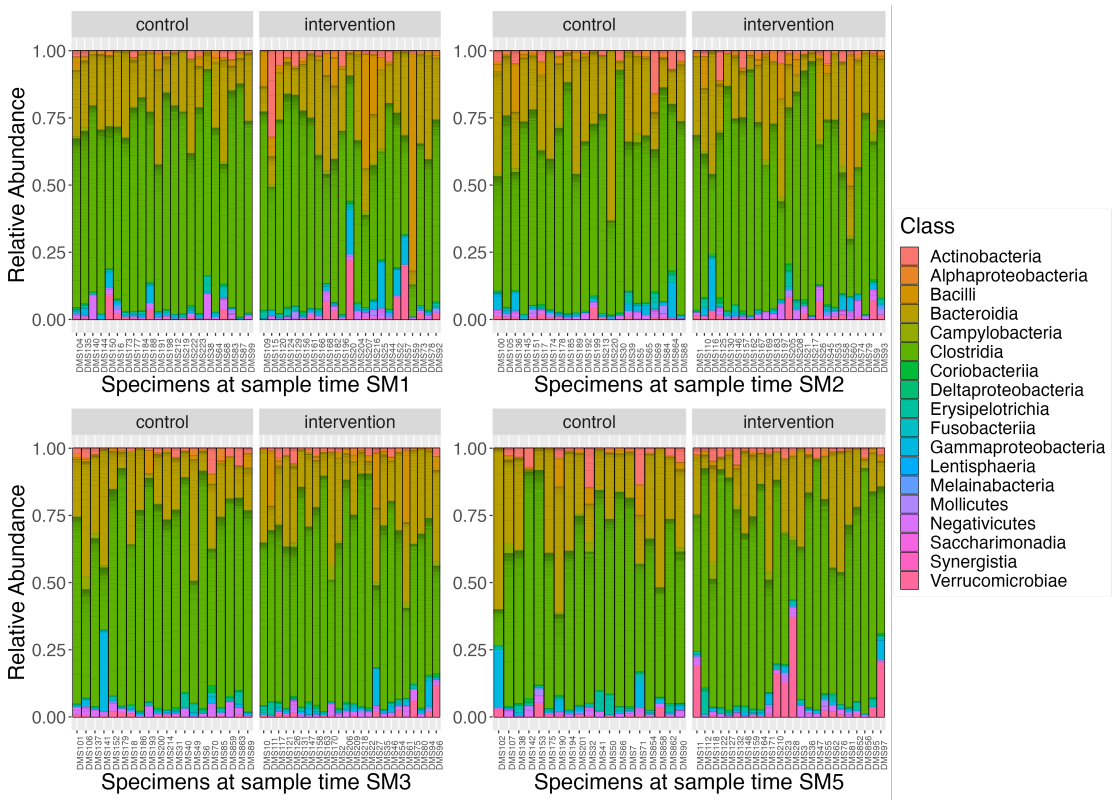


Figure 9: Relative abundance of specimens at the four-time points.

Figure 9 displays the distribution of relative abundance of specimens collected at the four time points coloured by the class levels of taxonomy taxa between control and intervention groups. We see that the class *Clostridia* is of high abundance in all time points. Also, it is clear that the relative abundance changes over time by class level which suggest that class-level taxa composition change over time.

We considered specimen-wise networks based on Jaccard distance using a maximum threshold of 0.8. Figure 10 illustrates the microbial composition changes over time for both treatment groups. For instance, in the intervention group, specimens 52 (participantID 92) and 57 (participantID 94) are grouped at the first time point, but both specimen 58 (participantID 94) and specimen 60 (participantID 95) are isolated at the second time point. In addition, while specimens 26 (participantID 80) and 169 (participantID 129) are grouped, following a group of specimen 27 (participantID 80), specimen 61 (participantID 95) and specimen 126

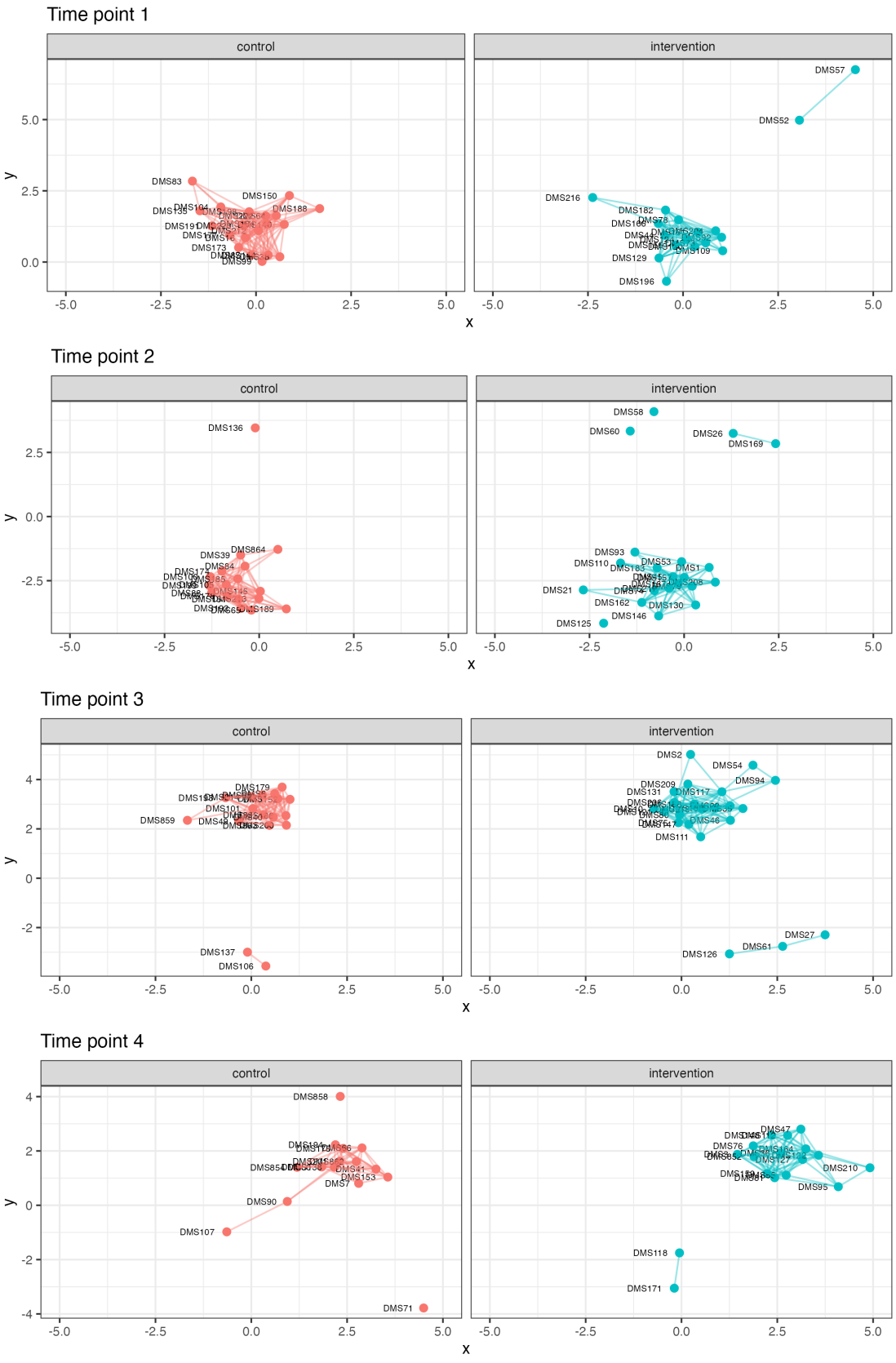


Figure 10: A network representation of microbial at the four-time points.

(participantID 115) at the thirst time point and finally specimen 118 (112) and specimen 171 (participantID 129) are grouping at the last time point. On the other hand, in the control group, specimen 136 (participantID 119) is isolated at the second time point, but specimen 106 (participantID 108) and specimen 126 (participantID 119) are grouped at the third time point. We conducted graph based tests (Jeganathan & Holmes, 2021) for the networks constructed at each time point for testing the null hypothesis that the microbial distribution of the two treatment groups are the same. The p-values of the test are 0.24, 0.908, 0.948 and 0.62 suggesting that there is not enough evidence to conclude microbial distributions are different between treatment groups at each time point.

We highlight that the exploratory tools and network-based hypothesis could not identify temporal changes in taxa.

3.3 Mixed Models for Taxa Abundance

To identify differentially abundance taxa in the experimental factors, we fitted NBMMs for each taxon. We considered removing additional taxa to access performance of mixed model and topic models. That is, we considered the top 150 taxa based on the total taxon abundances, resulting in 150 taxa in 181 specimens. Considering the treatment group (X_1) and time point at which the samples were collected (X_2) as fixed effects and random intercepts for participants ($\mathbf{Z}$), logarithm of the library size (N_d) as offsets, the NBMM for taxon abundance is given as follows.

$$\log(E(Y_v|X_1, X_2, \mathcal{B} = \boldsymbol{b})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathbf{Z}\boldsymbol{b} + log(N_d), \tag{3.1}$$

where X_1 = treatment group (binary; control = 0, intervention = 1),
X_2 = time point and

$$\mathbf{Z} = \begin{pmatrix} Participant_1 \\ Participant_2 \\ \vdots \\ Participant_s \end{pmatrix}^T = \begin{pmatrix} 1 & ... & 1 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & ... & 0 & 1 & ... & 1 & 0 & ... & 0 \\ \vdots & ... & \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots \\ 0 & ... & 0 & 0 & ... & 0 & 1 & ... & 1 \end{pmatrix}^T .$$

After fitting the NBMMs for each of 150 taxa, p values corresponding to the variable group (X_1) were adjusted using the Benjamini and Yekutieli (2001) method for controlling false discovery rate for correlated hypothesis, in order to identify the taxa that are differentially abundant between the control and intervention groups. Figure 11 shows the abundance and transformed abundance (described in Section 2.2) of the significant taxa. The transformed abundance of most of the significant taxa in the control group decreases in the second time point but increased in the third time point and again decreases in the last time point. On the other hand an opposite pattern can be observed for most of the significant taxa in the intervention group.

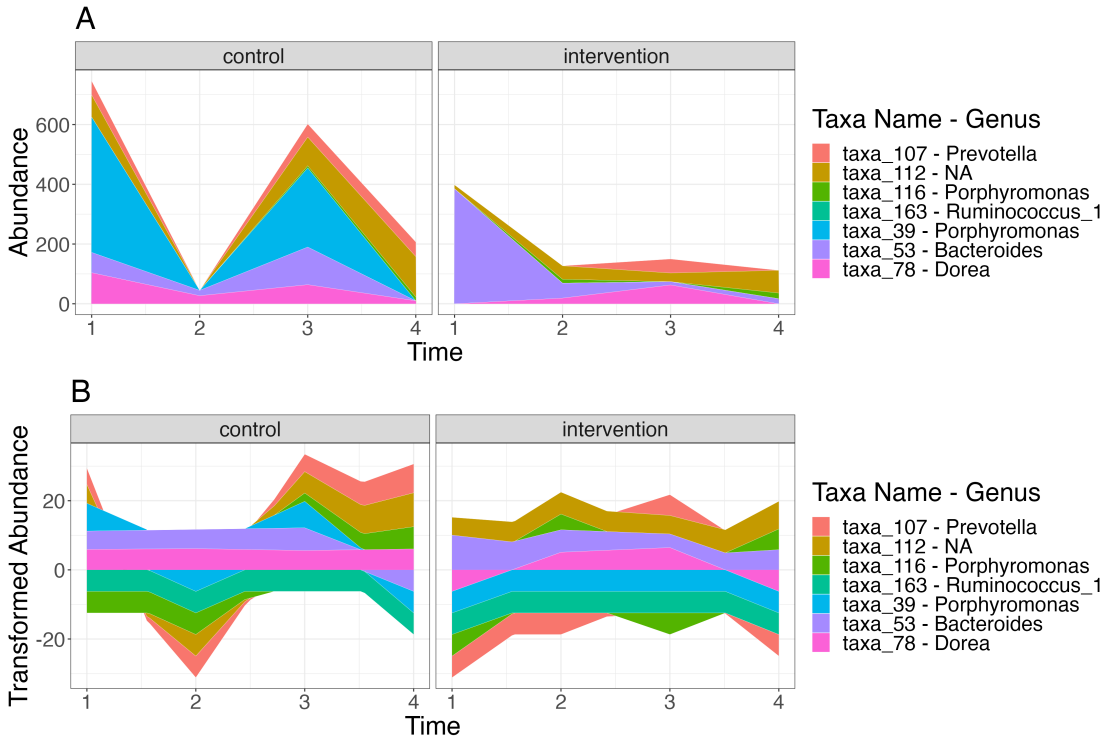


Figure 11: A: Abundance and B: Transformed abundance of significant taxa over time.

Note that Figure 11: (A) is hard to interpret because of library size effect and over-dispersion.

Chapter 4

Application of LDA and time-aligned LDA

This chapter presents the application of LDA and time-aligned LDA models on BHIP example data set. First, we present the implementation LDA models, followed by their corresponding results, and finally giving the results of the mixed model implementation with relevant conclusions.

4.1 LDA

This section focuses on the implementation of LDA topic model. We applied LDA on BHIP data after filtering the top 150 taxa as explained in Section 3.3. Hence a total of 150 taxa corresponding to 181 specimens were used in the application of LDA.

4.1.1 Identifying Number of Topics

The number of topics, K, should be pre-specified for LDA. We used recently introduced `alto` R package for this purpose (Fukuyama et al., 2023). The optimal

K was selected by conducting topic alignment across different LDA models using two methods; one that uses the specimen composition known as "Product alignment" and one that uses the topic composition known as "Transport alignment". We used three diagnostic scores; number of paths, coherence scores and refinement

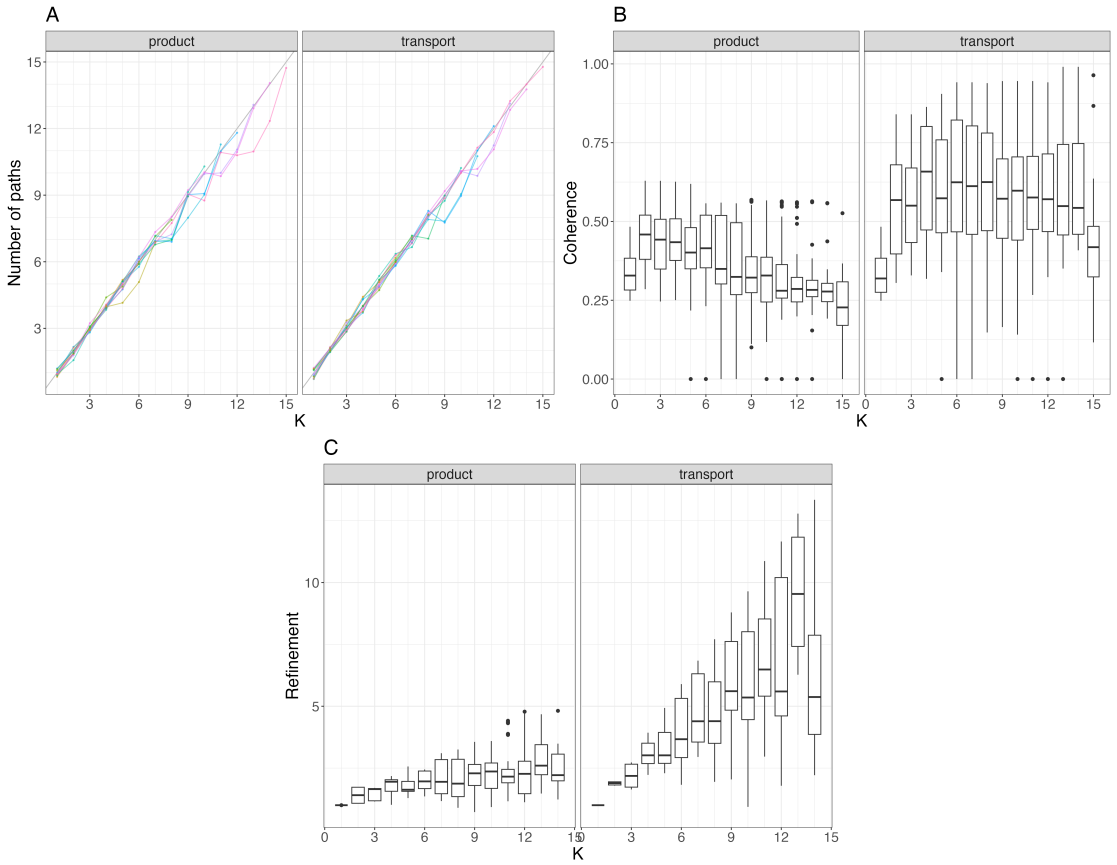


Figure 12: Alignment diagnostic scores; A: Number of paths, B: Coherence scores, C: Refinement scores. Colors denote the various alignments.

scores (Fukuyama et al., 2023) in deciding the number of topics. Topic alignment was done on LDA models that were fitted for different number of topics K ranging from 5 to 15 (Symul et al., 2023). The corresponding diagnostics plots are illustrated in Figure 12. We identified a deviation of the number of paths and a decrease in the coherence and refinement scores at $K = 5$ and $K = 8$ for product and transport methods, respectively. Therefore, the product method suggests $K = 4$ while the transport method suggests $K = 7$ as the optimal number of topics. For a simple model, we set $K = 4$. Figure 13 demonstrates the evolution

of LDA models as the number of topics K increase from 1 to 4.

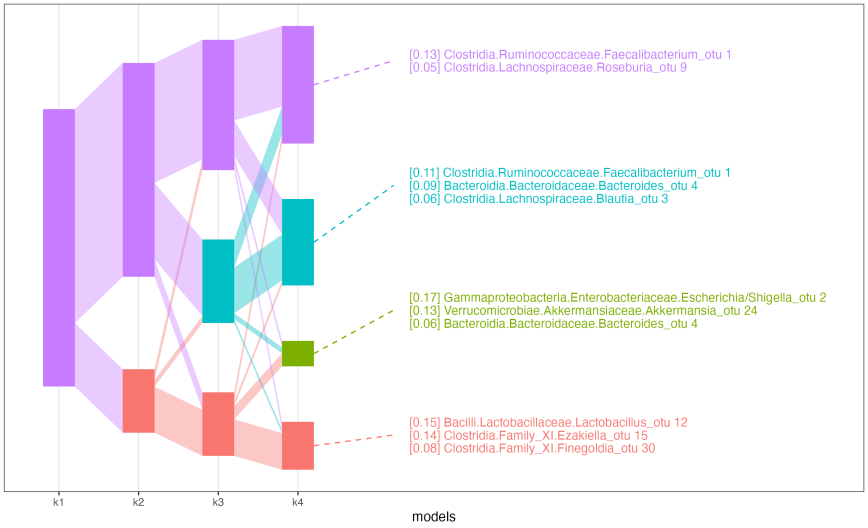


Figure 13: Alignment of different LDA models based on product method for chosen K, 1 to 4.

4.1.2 LDA Implementation

After identifying $K = 4$, we set the Dirichlet hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_4)$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{150})$ such that $\alpha_1 = ... = \alpha_4 = 0.8$ and $\gamma_1 = ... = \gamma_{150} = 0.8$. This parameter was set to a value less than 1 with the intention of generating sparse mixtures of taxa and topic proportions that are different and realistic in the posterior estimates. We use HMC-NUTs with 8000 iterations to draw samples from the posterior distribution of θ and B in (2.8). Further, we used first 4000 iterations as warm-up. We also checked the trace plots for some of the parameters to ensure the warm-up iterations.

We extracted the samples of θ and B in a three dimensional array. Denoting the number of iterations excluding warm-up runs as I, for D specimens the dimension of θ is given by $I \times D \times K = 4000 \times 181 \times 4$ and for V taxa the dimension of B is given by $I \times K \times V = 4000 \times 4 \times 150$. A representation of posterior estimates for topic proportion θ_d in specimen d and for taxa proportion B in topic k are illustrated in Tables 8 and 9.

Table 8: Posterior estimates of topic proportion for specimen d. θ_{d_i} is the i^{th} iteration estimate of θ for topic k of specimen d.

Specimen_d	Topic_1	...	Topic_K
iter 1	θ_{d1_1}	...	θ_{dK_1}
iter 2	θ_{d1_2}	...	θ_{dK_2}
$\vdots$	$\vdots$	...	$\vdots$
iter 4000	$\theta_{d1_{4000}}$	...	$\theta_{dK_{4000}}$

Table 9: Posterior estimates of taxa proportions for topic k. β_{kv_i} is the i^{th} iteration estimate of β for taxa v of topic k.

Topic_k	Taxa_1	...	Taxa_V
iter 1	β_{k1_1}	...	β_{kV_1}
iter 2	β_{k1_2}	...	β_{kV_2}
$\vdots$	$\vdots$	...	$\vdots$
iter 4000	$\beta_{k1_{4000}}$	...	$\beta_{kV_{4000}}$

4.1.3 LDA Model Diagnostics

The next step involves conducting model diagnostics. It is a vital step as it evaluates whether we draw samples from the posterior distribution and have sufficient effective samples for each parameter. The diagnostics provide directions for hyper-parameter tuning of HMC-NUTs and prior distributions.

Usually, the samples will be auto correlated within a Markov chain and treating samples as independent would underestimate the uncertainty associated with each parameter as it is inversely proportional to $\sqrt{N}$ where N is the sample size. Therefore, we use a diagnostic measure known as 'effective sample size' (ESS). ESS denoted by N_{eff} quantifies the effective number of samples generated in the Markov chain accounting for auto correlation.

Suppose $X_1, ..., X_N$ is sequence of Markov chain with N number of iterations. Then, the auto correlation for time lag t is defined as,

$$\rho_t = \frac{Cov(X_i, X_{i+t})}{\sqrt{Var(X_i)Var(X_{i+t})}}, \tag{4.1}$$

and the ESS of N samples generated by a process with auto correlations ρ_t is

defined as follows.

$$N_{eff} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}. \tag{4.2}$$

The ESS is considered as the number of iterations for independent draws, is lower than the number of iterations for correlated draws and is larger than the number of iterations for anti-correlated draws. HMC-NUTs can produce anti-correlated draws if the posterior is close to Gaussian with weak correlation (Stan et al., 2023).

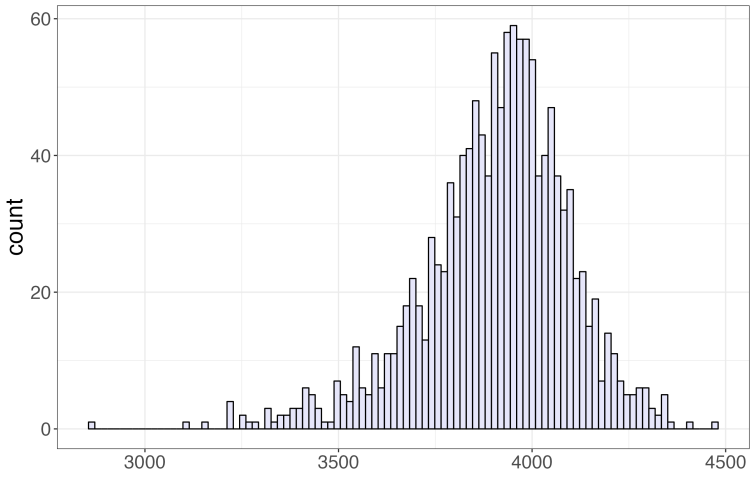


Figure 14: Effective sample size (ESS) for each parameter estimation.

Figure 14 demonstrates the distribution of ESS obtained for all parameters in HMC-NUTs. We observe that the ESS is highly dense around 4000 which is the number of iterations $I = 4000$, suggesting that the number of samples generated during the posterior simulation is sufficient such that the posterior sampling process is reliable, and no increment is needed in the number of iterations.

4.1.4 LDA Model Assessment

Further, abundance data were simulated using the posterior estimates as described in (2.5) and compared with the observed abundance data in order to assess the goodness-of-fit. Model assessment allows to evaluate model's reliability in cap-

turing the underlying latent variables under the model generative process.

For this purpose, we identified the maximum of simulated cell counts in each iteration and plot its distribution against the maximum of observed cell abundance in each taxon for several randomly selected taxa in the observed data. Figure 15 shows that most of observed maximum falls in the range of simulated maximums, suggesting a good-fitted model. However, for some taxa, the observed maximum lies in the tail of the simulated maximums, for which further investigation might be required. We may adjust the hyper-parameters for those taxa to increase the model performance.

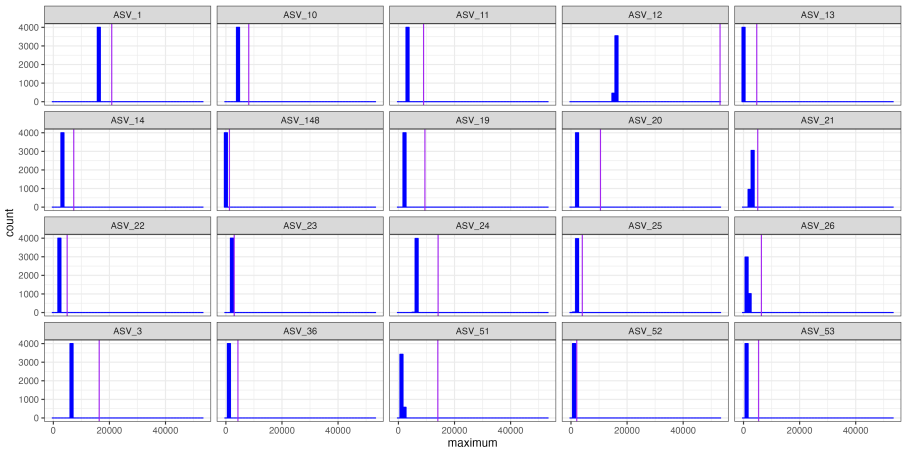


Figure 15: Model assessment for the LDA for several selected taxa. Blue histogram is the maximums from the simulated data. Purple vertical line is the observed maximum.

4.1.5 Estimated Topic and Taxa Distributions

The main interest in LDA is to obtain the estimates of topic distribution for each specimen θ_d and taxa distribution for each topic B. The estimates θ_d and B are obtained by taking the median of the posterior samples are illustrated in figure 16 and 17.

Figure 16 shows on average, the trend in topic proportions is similar between treatment groups for all topics, and there is a small change in Topics 1 and 2 over time. Figure 16: (A) shows that Topic 2 and 3 are dominating at all time

points regardless of the treatment groups. Interestingly, topic proportions are high at time point 1 and start to reduce over the time for Topic 1 in the intervention group, whereas the same topic starts at low and increase over time in the control group. Figure 17 illustrates the taxa composition of topics. For instance, Taxa 14 which belongs to the *Lactobacillus* genus dominates in Topic 1, and Taxa 39, 69, 79, 86, 101, 107, 113, 152, 153, 162 dominates in Topic 4 which mostly consists of *Porphyromonas* and *Prevotella* genera. Topic 2 and 3 are compounds from almost all of the taxa while dominated by *Bacteroides* and *Roseburia* genera. Interestingly, taxa 39 and 107 were found to be differentially abundant between the two treatment groups in Section 3.3, which are here found to be present in Topic 4. We use mixed models to infer whether topics are differentially abundant in Section 4.3.

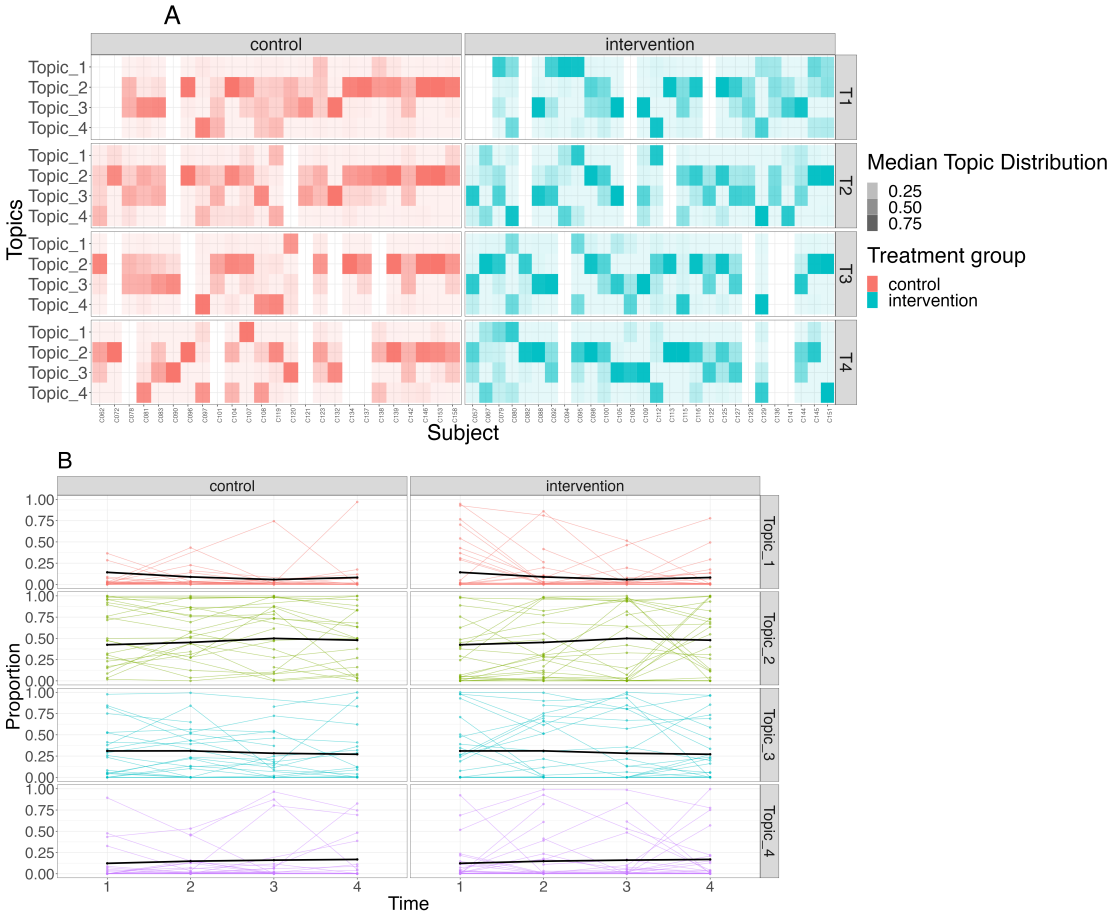


Figure 16: Estimated topic distribution in specimens obtained from LDA; A: Heat map of median topic proportions, B: Topic proportion of subjects over time.

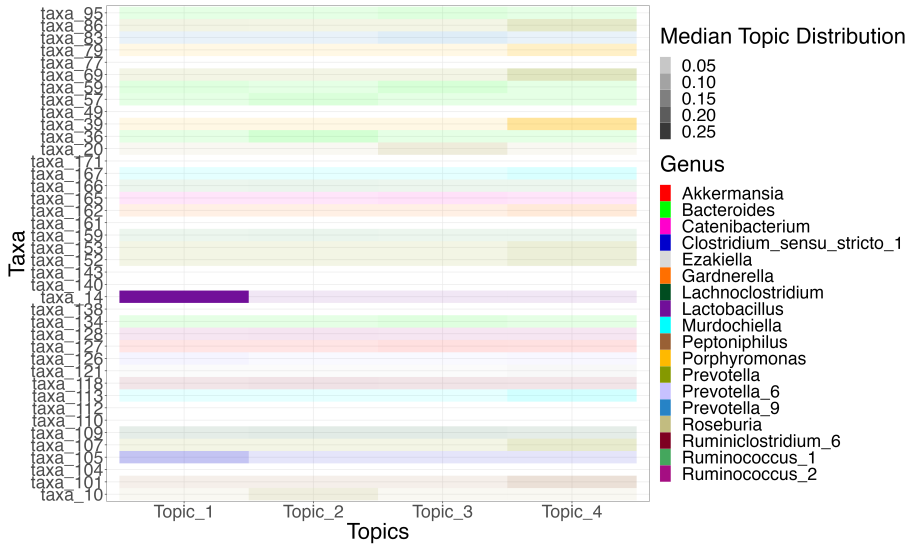


Figure 17: Estimated median taxa proportions in topics obtained from LDA.

4.2 Time-aligned LDA

This section focuses on the implementation of time-aligned LDA topic model. The same data with a total of 150 taxa in to 181 specimens were used in this context as well. As explained in Section 2.4.1, initially four cohorts corresponding to the four time points were generated using normalized exponential weights W given in (4.3), specifically using w_i as the sampling distribution when creating the i^{th} cohort C_i where $i = 1, 2, 3, 4$, which eventually resulted cohort 1 with 91 specimens, cohort 2 with 92 specimens, cohort 3 with 89 specimens and cohort 4 with 90 specimens.

$$W = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} 1 & 0.4324 & 0.3243 & 0.2432 \\ 0.3636 & 1 & 0.3636 & 0.2727 \\ 0.2727 & 0.3636 & 1 & 0.3636 \\ 0.2432 & 0.3243 & 0.4324 & 1 \end{pmatrix}. \tag{4.3}$$

4.2.1 Local LDA Implementation

After creating the four cohorts, local LDA models were fitted for each cohort independently. For comparative purposes, the parameters used in the LDA, namely; number of topics $K = 4$, Dirichlet hyper-parameters $\alpha_1 = ... = \alpha_4 = 0.8$ and $\gamma_1 = ... = \gamma_{150} = 0.8$ were set in the local LDA implementations. For each cohort, we used HMC-NUTs with 8000 iterations of which the first 4000 are considered as warm-up runs to draw posterior samples for θ_d and B.

We extracted four sets of estimates of θ_d and B in posterior samples in three dimensional arrays as mentioned in Section 4.1.2, corresponding to the 4 local LDA models. Denoting the number of specimens in the i^{th} cohort C_i as n_i, the dimension of θ corresponding to C_i is given by $I \times n_i \times K$ and for all the cohorts the dimension of B is given by $I \times K \times V = 4000 \times 4 \times 150$.

4.2.2 Model Diagnostics and Model Assessment

As for the LDA, we conducted model diagnostics using ESS and model assessment for the same selected taxa as the time-unaware LDA, for the 4 local LDAs. Figures 18 and 19 depict the distribution of ESS obtained for each parameter estimation and the distribution of the maximum of simulated cell counts in each iteration against the maximum of observed cell abundance in each taxon for the selected taxa, respectively. The ESS of all local LDAs are highly dense around 4000, therefore the number of samples generated during the posterior sampling are sufficient. Further, the observed maximum lie in the range of simulated maxima. Thus, the model is a good-fit.

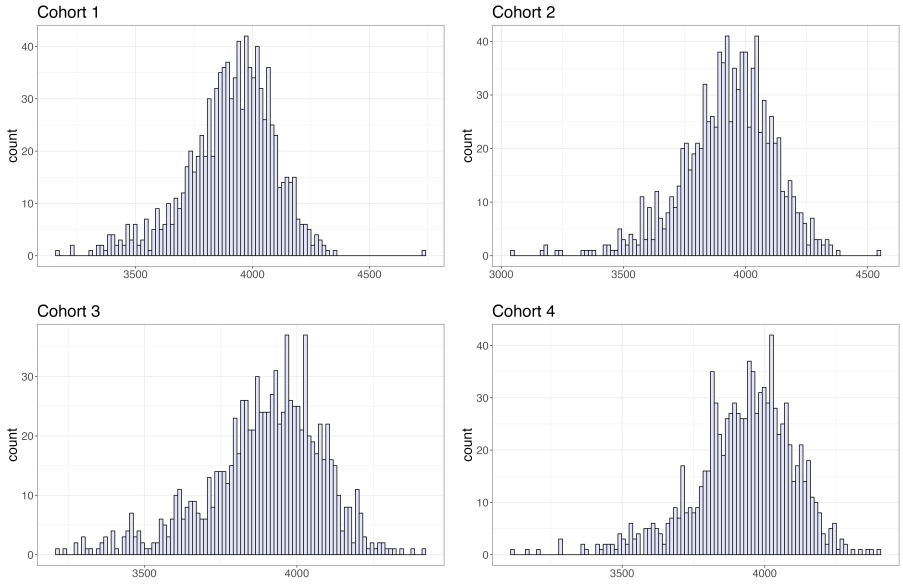


Figure 18: Effective sample size (ESS) for each parameter estimation of the 4 local LDA fitted for the 4 cohorts. Blue histogram is the maximums from the simulated data. Purple vertical line is the observed maximum.

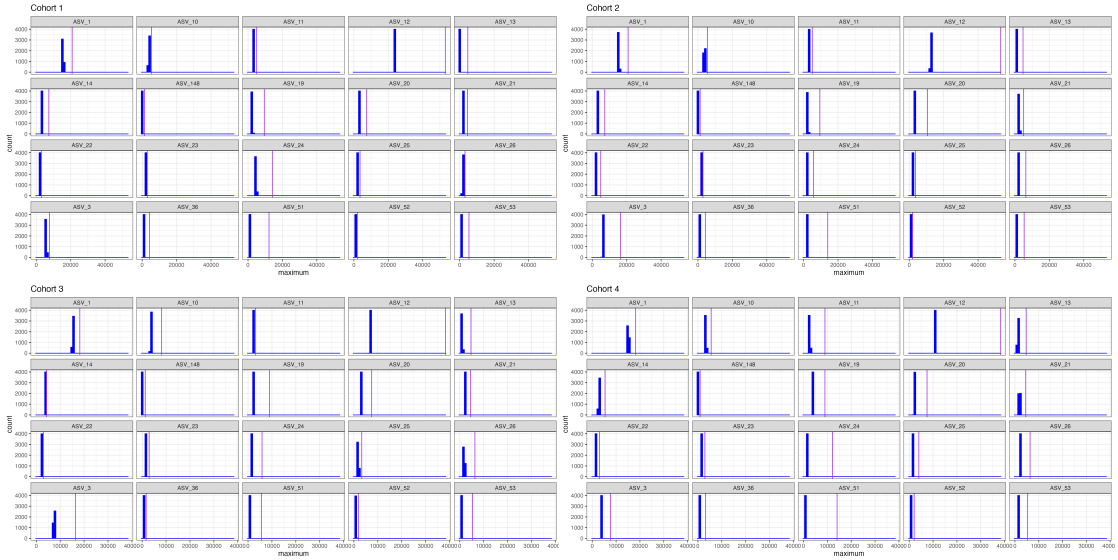


Figure 19: Model assessment for the local LDA for the selected taxa.

4.2.3 Estimated Topic and Taxa Distributions

Once the local LDAs were constructed, the corresponding estimates of topic proportions θ_d in specimens d and taxa distributions in topics B were obtained by taking the median of the posterior samples for local LDA separately. Then we

aligned the topics of local LDA considering cohort 1 as a reference. Next, we obtain a time-aligned topic distribution in specimens by merging the cohort results. This process is explained in detail in Section 2.4.2.

Figure 22 illustrates the estimated median topic proportions in specimens. As similar to LDA, there is no difference in the average topic proportion over time between the control and intervention groups for all the topics. However, now we can clearly observe a decreasing trend in Topic 1, increasing trend in Topic 4 and fluctuating trend in Topic 2. Further, topic proportions are high at first time point and reduce over the time for Topic 1 in the intervention group. Topic 4 proportions are low at first time point and increase over the time in the control group. Moreover, the changes in taxa composition in each topic are illustrated in Figure 21. For instance, in Topic 4, Taxa 14 and 105 which belong to *Lactobacillus* and *Clostridium* genera dominate at time point 1, Taxa 20 and 36 in the *Roseburia* and *Bacteroides* genera dominates at time point 2, Taxa 20 in *Roseburia* genus and Taxa 14 in *Lactobacillus* genus dominates at time points 3 and 4. More interestingly, Taxa 39, 53, 107, 112, 116 and 168 that were found to be differentially abundance between the treatment groups are present in the topics identified in tme-aligned LDA model.

Next, we find the significantly differential topics in experimental factors using mixed models.

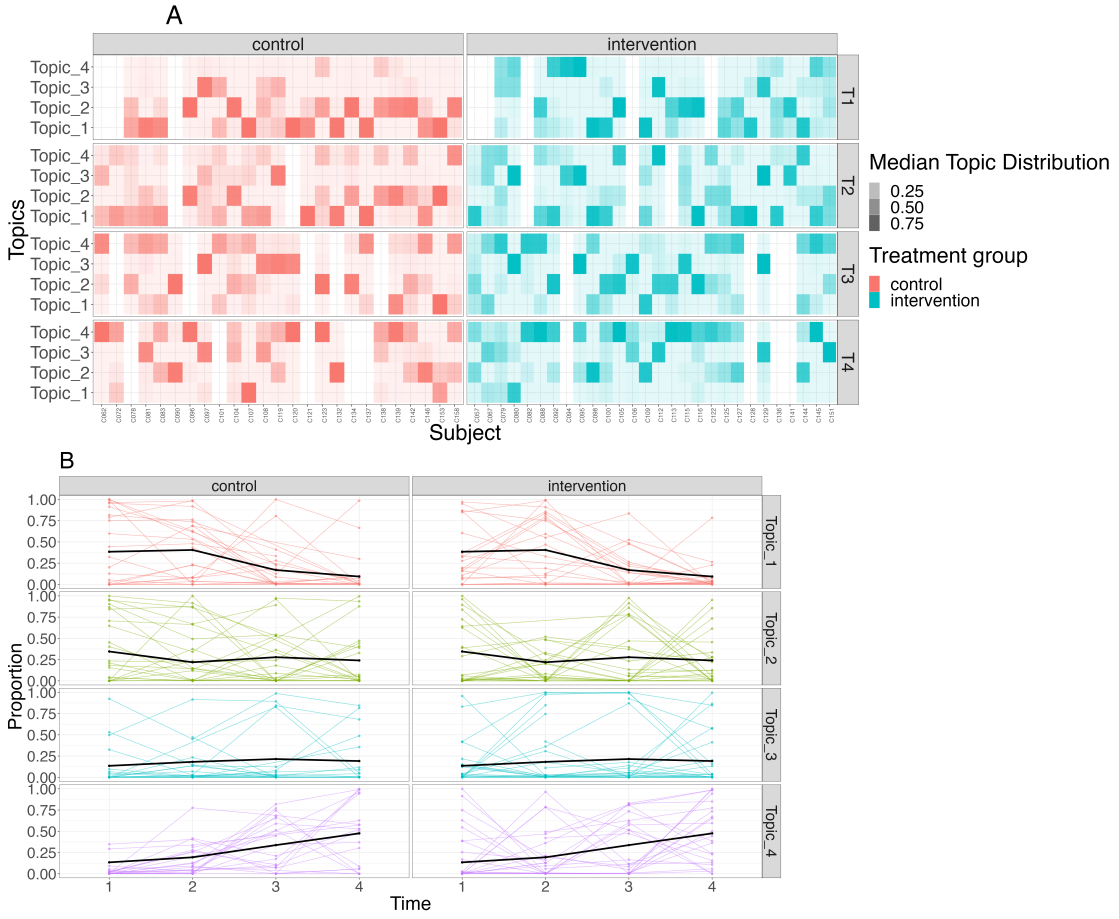
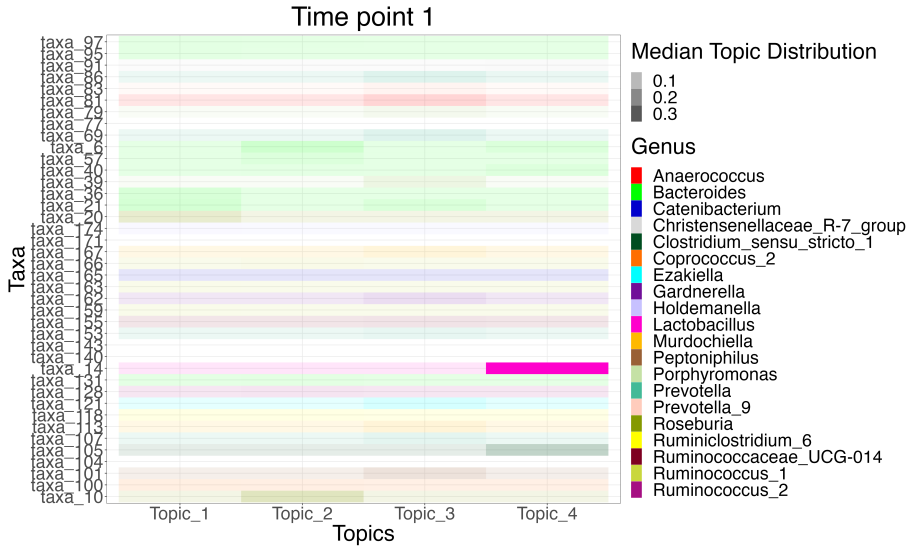


Figure 20: Estimated topic distribution in specimens obtained from time-aligned LDA; A: Heat map of median topic proportions, B: Topic proportion of subjects over time.



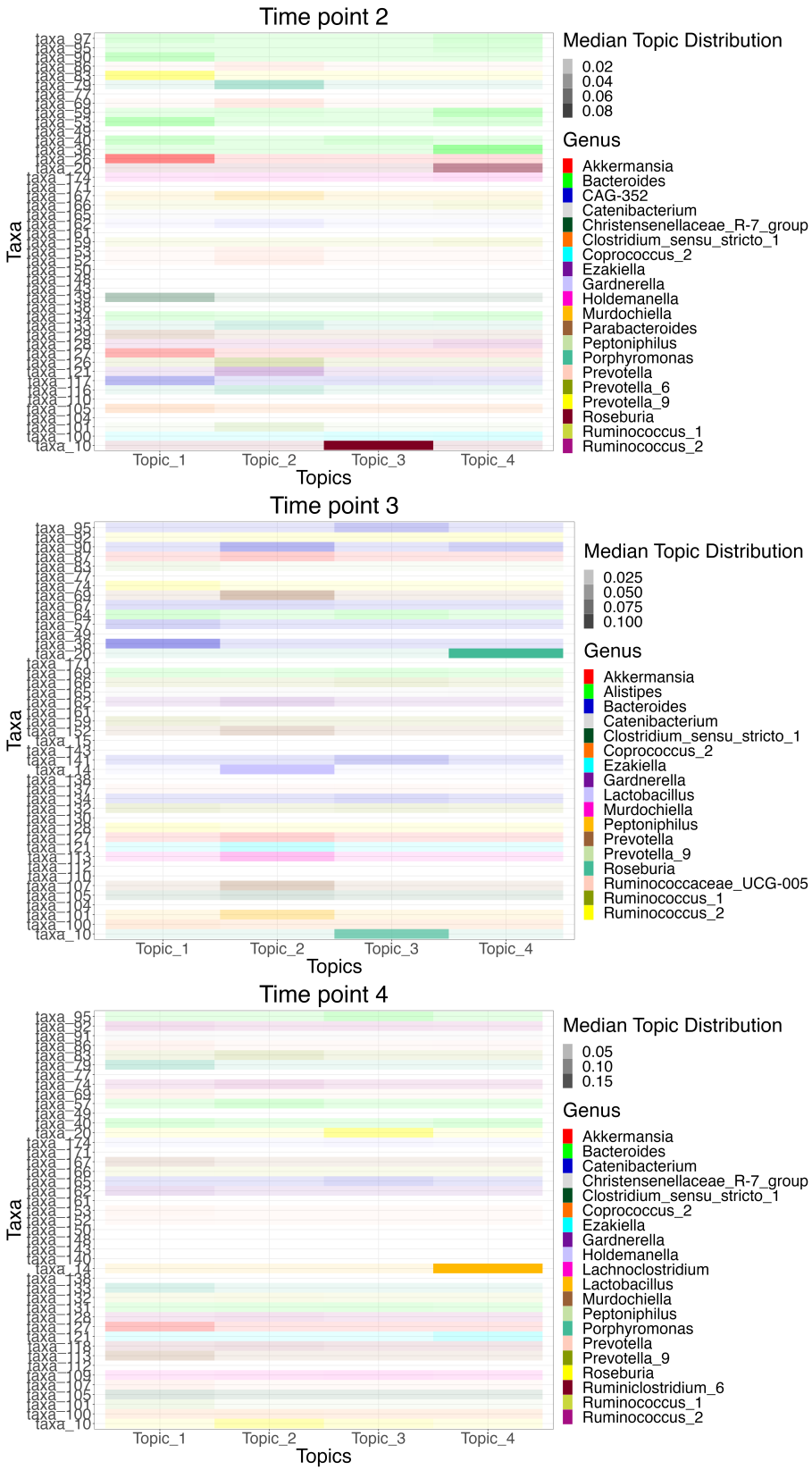


Figure 21: Estimated taxa distribution in topics obtained from time-aligned LDA.

4.3 Mixed Models for Topic Proportions

LMMs were fitted for the logarithm of topic proportion of each topic obtained from both the time-unaware and time-aligned LDA models. In (4.4) the treatment group (X_1) and time point at which the samples were collected (X_2) were considered as fixed effects and random intercepts for the participants ($\mathbf{Z}$) were considered.

$$\log(P_k|X_1, X_2 d, \mathcal{B} = \boldsymbol{b}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathbf{Z}\boldsymbol{b}, \tag{4.4}$$

where X_1 = treatment group (binary; control = 0, intervention = 1),

X_2 = time point,

$$\mathbf{Z} = \begin{pmatrix} Participant_1 \\ Participant_2 \\ \vdots \\ Participant_s \end{pmatrix}^T = \begin{pmatrix} 1 & ... & 1 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & ... & 0 & 1 & ... & 1 & 0 & ... & 0 \\ \vdots & ... & \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots \\ 0 & ... & 0 & 0 & ... & 0 & 1 & ... & 1 \end{pmatrix}^T.$$

The results of the LMM's are given in Tables 10 and 11. In time-unaware LDA, there is a significant difference between the treatment groups for Topic 2 and a significant difference in at least one of the time points for Topic 1. In contrast, in the time-aligned LDA, there is not enough evidence to find significant topic differences between the treatment groups. In contrast to only one topic changes significantly over time in the time-unaware LDA, time-aligned LDA found Topic 1 and 4 significantly different at least one of the time points.

Next, we compared how well the topics and temporal changes are captured by the time-aligned LDA with LDA. We carried a holdout specimen prediction by holding out one specimen at a time, fitting LMM on the rest of the specimens and compute the root squared error of the predictions of each topic. Figure 22 and Table 12 shows the root squared errors of the holdout predictions is similar between the time-unaware and time-aligned LDA in topics other that the one that has a significant group effect in the time-unaware LDA. However, the root squared

errors in the time-aligned LDA are larger than the time-unaware LDA.

Table 10: Linear mixed effect model results for LDA.
(In each row block, first row: regression coefficient, second row: (standard error of the regression coefficient), third row: p value, * indicates the p values significant at 5% significance).

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
Intercept	-4.6384	-1.4766	-3.3432	-5.5710
	(0.7203)	(0.5787)	(0.7986)	(0.7021)
	2.08e-09*	0.0123*	6.31e-05*	1.37e-12*
$\text{group}_{Intervention}$	0.0439	-1.5860	-0.3051	1.0401
	(0.7211)	(0.6590)	(0.9201)	(0.7499)
	0.9517	0.0198*	0.742	0.171
time point	-0.4157	0.1011	-0.1183	0.0461
	(0.2022)	(0.1329)	(0.1787)	(0.1811)
	0.0416*	0.4484	0.509	0.800

Table 11: Linear mixed effect model results for time-aligned LDA.
(In each row block, first row: regression coefficient, second row: (standard error of the regression coefficient), third row: p value, * indicates the p values significant at 5% significance).

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
Intercept	-1.8412	-3.3377	-5.7523	-6.6565
	(0.7514)	(0.7339)	(0.7429)	(0.7026)
	0.0154*	1.05e-05*	1.85e-12*	<2e-16*
$\text{group}_{Intervention}$	-0.1979	-0.7310	0.9224	0.2110
	(0.6958)	(0.6150)	(0.7317)	(0.5654)
	0.7772	0.240	0.213	0.711
time point	-0.8360	-0.0584	0.1576	1.1690
	(0.2273)	(0.2379)	(0.2122)	(0.2329)
	0.0003*	0.806	0.459	1.49e-06*

Table 12: Average of the root squared errors of holdout sample predictions.

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
LDA	2.69	1.40	2.20	2.38
Time-aligned LDA	3.03	3.26	2.75	2.94

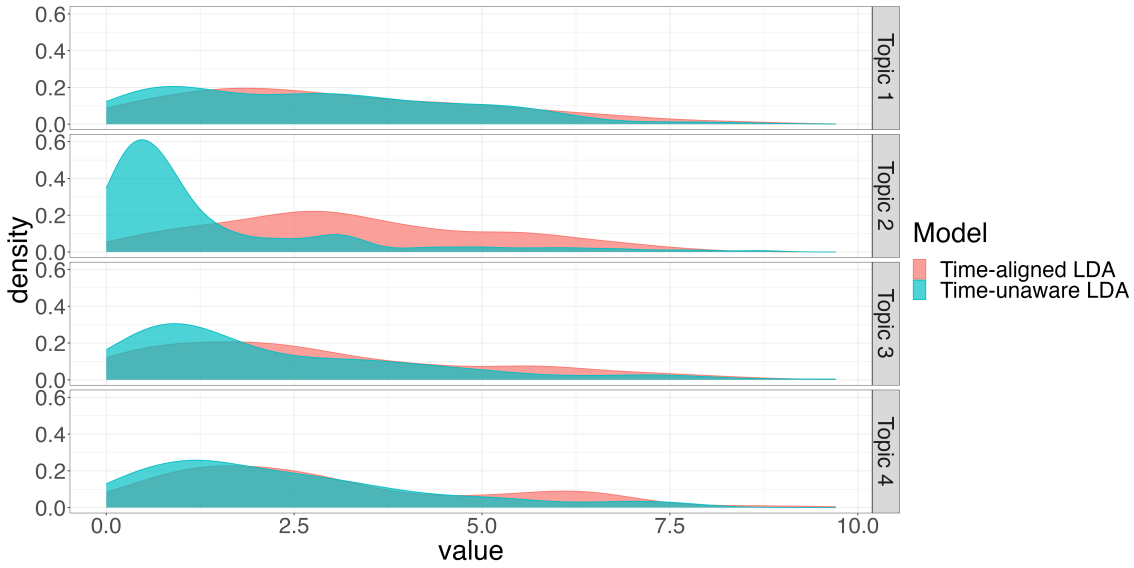


Figure 22: Distributions of root squared errors of holdout specimen predictions.

Chapter 5

Discussion

This chapter provides a discussion based on the literature review and the findings of the research. Further, it concludes the thesis with limitations of the study and suggestions for future research.

Microbial data are dynamic and driven by the interactions of taxa between them and with the host, which has lead to increase in studying microbial data longitudinally. More importantly, identifying the temporal dynamics of microbial communities instead of individual taxa is insightful for understanding functionality of taxa co-exist. With respect to identifying microbial communities, the implementation of probabilistic LDA model proposed by Sankaran and Holmes (2019) is popular as it is applicable for multivariate, high dimensional, sparse data while allowing for mixed membership for specimens in the clusters. Thus, this study proposes a time-aligned LDA for longitudinal microbiome data with the inspiration obtained from Wang et al. (2021) for identifying temporal changes in microbial communities.

The proposed time-aligned LDA involves, creating temporal smoothed cohorts for every time point, implementing LDA for each of those cohorts independently, followed by aligning cohort results. The parameters; number of topics $K = 4$,

Dirichlet hyper-parameters $\boldsymbol{\alpha} = \boldsymbol{\gamma} = 0.8$ were set in all local LDA implementations. HMC-NUTs was used for posterior sampling with 8000 iterations while using first 4000 as warm-up runs. Using 4000 iterations for posterior sampling provided a sufficient sample size such that the posterior sampling is reliable. The median of posterior samples were taken as the estimated topic proportions in specimens and taxa proportions in topics. Further, the taxa composition in all the chosen four topics change at each time point which suggest there is a temporal change in those four microbial communities over time.

Pregnancy is one of the most suitable events to observe changes in human microbiome as a result of not just physical but also hormonal, immunoloigical and metabolic changes that a pregnant woman experience in the course of body adapting for fetal growth and development (Atkinson et al., 2022, Symul et al., 2023). Hence, the proposed method was applied on Be Healthy In Pregnancy (BHIP) Atkinson et al., 2022 data set which contains gut microbiome specimens corresponding to 63 pregnant women in southern Ontario, Canada, who were randomly allocated to either the intervention group which provided nutrition, exercise and counselling or the control group. Data are available for 4 time points, corresponding to each trimester and one time after the pregnancy.

According to the results, estimated topic proportions in specimens had a significant group for one topic in the LDA, but for no topics in the time-aligned LDA. This may be due to small group differences in topic proportions obtained in the time-aligned LDA which might not captured with the given sample size. Further, estimated topic proportions in specimens had a significant time effect at least in one of the time points for one topic in the LDA, but for two topics in the time-aligned LDA. Even though there is no significant impact of the treatment, the microbial composition is expected to change over time and the time-aligned LDA and mixed models identified more topics that are temporally changing. This might be due to capturing microbial community changes locally in time-aligned

LDA. More importantly, the proposed time-aligned LDA provide information on how the taxa composition in topics change at each time point, which cannot be derived from the LDA. However, the holdout sample predictions errors are higher for the proposed time-aligned LDA. This might be because we assumed all taxa are present in all the time points, even though we see there are some taxa not present at some time points. To overcome this issue, we might need to consider sparsity imposed on the time-aligned LDA. Further, there is subject and specimen heterogeneity captured in microbial communities when we use time-aligned LDA which is not captured over time in the mixed models, which might also be a reason for the higher prediction errors in the time-aligned LDA. Regardless, the proposed time-aligned LDA was able to identify significant microbial community changes over time in terms of topic proportions as well as the taxa composition changes of those topics over time, which eventually provides more useful information.

5.1 Study Limitations and Future Research

This study only focused on top 150 taxa because of the memory restrictions in applying all taxa in the local compute, but all the taxa can be considered with the help of high-performance computing.

Next, in the implementation of LDA, a hyper-parameter tuning was not done due to limited time. Thus, some of the resulted topics did not capture dominating taxa in them. We can consider tuning of the LDA hyper-parameters in topics to sparse taxa proportions. Or one can add a prior over the hyper-parameters in the LDA and letting the MCMC algorithm mix among them rather than performing hyper-parameter tuning. This approach is naturally Bayesian and can be more efficient.

Also, in the implementation of the proposed time-aligned LDA, we used the same prior distribution for all cohorts. Instead, either conducting a separate hyper-

parameter tuning for each local LDA or seeding the next LDA with estimates of the current LDA can be done (Wang et al., 2021).

One can apply the proposed method to a highly dynamic microbiome ecosystem such as vaginal microbiome which changes in pregnant women (Symul et al., 2023) to investigate the temporal changes in microbial communities.

In the future, we need to perform a simulation study in order to investigate the properties of the time-aligned LDA.

Appendix

The code to reproduce the results and figures can be found at https://github.com/IshankaRandini/Master-Thesis-Sirikkathuge-Fernando.git.

References

Alshawaqfeh, M., Serpedin, E., & Younes, A. B. (2017). Inferring microbial interaction networks from metagenomic data using SgLV-EKF algorithm. *BMC Genomics*, *18*, 1–12.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, *35*(3/4), 246–254.

Atkinson, S. A., Maran, A., Dempsey, K., Perreault, M., Vanniyasingam, T., Phillips, S. M., Hutton, E. K., Mottola, M. F., Wahoush, O., Xie, F., et al. (2022). Be healthy in pregnancy (BHIP): A randomized controlled trial of nutrition and exercise intervention from early pregnancy to achieve recommended gestational weight gain. *Nutrients*, *14*(4), 810.

Baksi, K. D., Kuntal, B. K., & Mande, S. S. (2018). TIME: A web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Frontiers in Microbiology*, *9*, 36.

Benincà, E., Pinto, S., Cazelles, B., Fuentes, S., Shetty, S., & Bogaards, J. A. (2023). Wavelet clustering analysis as a tool for characterizing community structure in the human microbiome. *Scientific Reports*, *13*(1), 8042.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bodein, A., Chapleur, O., Droit, A., & Lê Cao, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Frontiers in Genetics*, *10*, 963.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M. L., Liu, Q., et al. (2016). MDSINE: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biology*, *17*, 1–17.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583.

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics & Informatics*, *17*(1).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610.

Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., & Weitz, J. S. (2020). A primer for microbiome time-series analysis. *Frontiers in Genetics*, *11*, 310.

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998.

Fisher, C. K., & Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PloS One*, *9*(7), e102451.

Fukuyama, J., Sankaran, K., & Symul, L. (2023). Multiscale analysis of count data through topic alignment. *Biostatistics*, *24*(4), 1045–1065.

Gerber, G. K., Onderdonk, A. B., & Bry, L. (2012). Inferring dynamic signatures of microbes in complex host ecosystems. *Plos Computational Biology*.

Gibson, T., & Gerber, G. (2018). Robust and scalable models of microbiome dynamics. *International Conference on Machine Learning*, 1763–1772.

Guo, J., Gabry, J., Goodrich, B., & Weber, S. (2020). Package 'rstan'. *URL https://cran. r—project. org/web/packages/rstan/(2020)*.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PloS One*, *7*(2), e30126.

Holmes, S., & Huber, W. (2018). *Modern statistics for modern biology*. Cambridge University Press.

Jeganathan, P., & Holmes, S. P. (2021). A statistical perspective on the challenges in molecular microbial biology. *Journal of Agricultural, Biological and Environmental Statistics*, *26*(2), 131–160.

Joseph, N., Paulson, C., Corrada Bravo, H., & Pop, M. (2013). Robust methods for differential abundance analysis in marker gene surveys. *Nature Methods*, *10*, 1200–1202.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kodikara, S., Ellul, S., & Lê Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*, *23*(4), 1–18.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., & Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, *7*, 1–14.

Mao, J., Chen, Y., & Ma, L. (2020). Bayesian graphical compositional regression for microbiome data. *Journal of the American Statistical Association*, *115*(530), 610–624.

McGeachie, M. J., Chang, H.-H., & Weiss, S. T. (2014). CGBayesNets: Conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data. *PLoS Computational Biology*, *10*(6), e1003676.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS computational biology*, *10*(4), e1003531.

Micah, H., Claire, F.-L., Rob, K., Gordon Jeffrey, I., et al. (2007). The human microbiome project: Exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804–810.

Nishio, M., & Arakawa, A. (2019). Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution*, *51*, 1–12.

Park, S.-Y., Ufondu, A., Lee, K., & Jayaraman, A. (2020). Emerging computational tools and models for studying gut microbiota composition and function. *Current Opinion in Biotechnology*, *66*, 301–311.

Rouyer, T., Fromentin, J.-M., Stenseth, N. C., & Cazelles, B. (2008). Analysing multiple time series and extending significance testing in wavelet analysis. *Marine Ecology Progress Series*, *359*, 11–23.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., & Narasimhan, G. (2021). Dynamic Bayesian networks for integrating multi-omics time series microbiome data. *Msystems*, *6*(2), 10–1128.

Sankaran, K., & Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostatistics*, *20*(4), 599–614.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Stan, D. T., et al. (2023). Rstan: The R interface to Stan. R package version 2.33.

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., & Xavier, J. B. (2013). Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Computational Biology*, *9*(12), e1003388.

Symul, L., Jeganathan, P., Costello, E. K., France, M., Bloom, S. M., Kwon, D. S., Ravel, J., Relman, D. A., & Holmes, S. (2023). Sub-communities of the vaginal microbiota in pregnant and non-pregnant women. *Proceedings of the Royal Society B*, *290*(2011), 20231461.

Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. (2021). A geometry-driven longitudinal topic model. *Harvard Data Science Review*.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., & Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, *9*, 1683.