Advances in Bayesian Hierarchical Modeling with Tree-based Methods

by

Jialiang Mao

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Advisor

_____
David Dunson

_____
Mike West

_____
Yongtao Guan

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2020

Abstract

Advances in Bayesian Hierarchical Modeling with Tree-based Methods

by

Jialiang Mao

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Advisor

_____
David Dunson

_____
Mike West

_____
Yongtao Guan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2020

Abstract

Developing flexible tools that apply to datasets with large size and complex structure while providing interpretable outputs is a major goal of modern statistical modeling. A family of models that are especially suitable for this task is the Pólya tree type models. Following a divide-and-conquer strategy, these tree-based methods transform the original task into a series of tasks that are smaller in size and easier to solve while their nonparametric nature guarantees the modeling flexibility to cope with datasets with a complex structure. In this work, we develop three novel tree-based methods that tackle different challenges in Bayesian hierarchical modeling. Our first two methods are designed specifically for the amplicon sequencing data in microbiome studies, which consists of high dimensional counts with a complex, domain-specific covariate structure and exhibits large cross-sample variations. These features limit the performance of generic statistical tools and require special modeling considerations. Both methods inherit the flexibility and computation efficiency from the general tree-based methods and directly utilize the domain knowledge to help infer the complex dependency structure among different microbiome categories by bringing the phylogenetic tree into the modeling framework. An important task in microbiome research is to compare the composition of the microbial community of groups of subjects. We first propose a model for this classic two-sample problem in the microbiome context by transforming the original problem into a multiple testing problem, with a series of tests defined at the internal nodes of the phylogenetic

tree. To improve the power of the test, we use a graphical model to allow information sharing among the tests. A regression-type adjustment is also considered to reduce the chance of false discovery. Next, we introduce a model-based clustering method for the microbiome count data with a Dirichlet process mixtures setup. The phylogenetic tree is used for constructing the mixture kernels to offer a flexible covariate structure. To improve the ability to detect clusters determined not only by the dominating microbiome categories, a subroutine is introduced in the clustering procedure that selects a subset of internal nodes of the tree which are relevant for clustering. This subroutine is also important in avoiding potential overfitting. Our third contribution proposes a framework for causal inference through Bayesian recursive partitioning that allows joint modeling of the covariate balancing and the potential outcome. With a retrospective perspective, we model the covariates and the outcome conditioning on the treatment assignment status. For the challenging multivariate covariate modeling, we adopt a flexible nonparametric prior that focuses on the relation of the covariate distributions under the two treatment groups, while integrating out other aspects of these distributions that are irrelevant for estimating the causal effect.

To my family

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathcal{T}$	A rooted full binary tree
$\mathcal{L}(\mathcal{T})$	The set of leaves of $\mathcal{T}$
$\mathcal{I}(\mathcal{T})$	The set of internal nodes of $\mathcal{T}$
$\mathcal{P}$	A tree-induced partition
A_l, A_r	The left and right child of a node $A \in \mathcal{I}(\mathcal{T})$

Abbreviations

AG	The American Gut Project
ASV	Amplicon Sequence Variants
ATE	Average Treatment Effect
ATT	Average Treatment Effect for the Treated
BCR	Bayesian Compositional Regression
BGCR	Bayesian Graphical Compositional Regression
CATE	Conditional Average Treatment Effect
CDR	Covariate Density Ratio
DM	Dirichlet Multinomial
DMM	Dirichlet Multinomial Mixtures
DTM	Dirichlet-tree Multinomial
DTMM	Dirichlet-tree Multinomial Mixtures

IBD	Inflammatory Bowel Disease
LNM	Logistic-normal Multinomial
PJAP	Posterior Joint Alternative Probability
PMAP	Posterior Marginal Alternative Probability
OPT	Optional Pólya Tree
OTU	Operational Taxonomic Units
PT	Pólya Tree
RBD	Randomized Block Design
RDDP	Recursive Dimensionwise Dyadic Partitioning
SI	Strong Ignorability
SUTVA	Stable Unit Treatment Value Assumption

Acknowledgements

First I want to express my sincere gratitude to my advisor Professor Li Ma, for the continuous support and guidance throughout my PhD adventure. Li is simply the best advisor I could ever imagine. He cares not only about my research, but also about my career development and well-being. His insights and enthusiasm in statistical research inspired me to become a better researcher. His kindness and consideration helped me become a better person.

I am grateful to my PhD prelim and dissertation committee members Professor David Dunson, Professor Mike West and Professor Yongtao Guan, for their instant feedback and invaluable suggestions on my dissertation research.

I am incredibly thankful to Duke FORGE for generously supporting my PhD research. I sincerely thank Professor Nazema Siddiqui, Professor Amy Herring and Professor Fan Li for helping me apply the Duke FORGE fellowship. I am much obliged to the National Science Foundation. My research at Duke is partly supported by the NSF Statistics Program grants DMS-1612889 and DMS-1749789.

I would like to thank all faculty members and staff at the Department of Statistical Science at Duke University. I am particularly thankful to Professor James Berger, Professor Alan Gelfand, Professor Surya Tokdar, Professor Peter Hoff, Professor Mine Cetinkaya-Rundel, Lori Rauch, Nicole Scott and Karen Whitesell.

I am indebted to all my friends at Duke University and LinkedIn Corporation. I want to thank my intern advisor Iavor Bojinov, for his help and advice. Special

thanks to Bai Li for being my incredible roommate, to Xi Chen, Ye Wang, Lu Wang, Shuxi Zeng, Jiurui Tang and Sheng Jiang for the meals, and to Xu Chen and Yuan Gao for the talks, the soccer games, and the beers. I also wish to thank my excellent classmates—Isaac Lavine, Kelly Moran, Felipe Ossa Monge, Yuhan Chen, Shai Gorsky, and Zach White, as well as everyone in Professor Li Ma's research group.

Finally, I want to express my deepest gratitude to my parents Gengsheng Mao and Yanjv Wang, and to my fiancé Can Feng. Without their unconditional and unlimited love and support, this long journey would be impossible. I dedicate this work to them.

1

Introduction

Nowadays, increasingly massive datasets become available by virtue of the evolving technologies to collect, preprocess, and store the data. Developing flexible tools that apply to such datasets while providing interpretable outputs is a major goal of modern statistical modeling. The massiveness of modern data is at least twofold—(i) the size of the dataset can be tremendous, and (ii) the structure of variables in the dataset can be sophisticated. These two facets pose seemingly irreconcilable modeling challenges: the complex data structure requires models to be flexible with a large number of parameters, while the large size of the data prefers simpler models with an easy computational solution. Besides its massiveness, a modern dataset is usually domain-specific, reflected in its complicated structure. Therefore, available domain knowledge can and should be incorporated into the modeling framework to produce interpretable analyses within the specific scientific domain.

Bayesian tree-based methods form an important family in the standard statistical toolkit. Some examples in this family include Bayesian classification and regression trees (BCART) (Chipman et al., 1998; Denison et al., 1998; Wu et al., 2007) and Bayesian additive regression trees (BART) (Chipman et al., 2010). A specific type

of models in this family—the so-called Pólya tree type models (Lavine et al., 1992, 1994; Paddock et al., 2003; Hanson, 2006; Wong et al., 2010)—is of special interests in modeling modern massive datasets. In this thesis, we investigate some members in this family with applications to the microbiome data, which has an inherent tree structure. These tree-based models are of interest for three reasons. Computationally, they offer a divide-and-conquer framework that breaks down the computation burdens into relatively small pieces which can be processed in parallel. From a modeling perspective, the nonparametric nature of these methods guarantees their flexibility while the tree structure ensures a natural way of reporting and interpreting the analysis results. Additionally, the way these methods construct coarse-to-fine resolutions and decompose empirical evidence to perform inference makes them ideal for detecting and reporting local signals that are otherwise likely to be submerged by the tides of noises.

In Chapter 2, we introduce the human microbiome and the microbiome sequencing data. Understanding the human microbiome is important for learning various aspects of our physiology. The maturity of next-generation sequencing technologies (Goodwin et al., 2016) together with the advance in preprocessing pipelines such as DADA2 (Callahan et al., 2016) give birth to large-scale microbiome sequencing datasets of high quality. A microbiome dataset contains counts of different microbes within a set of samples and provides a snapshot of their underlying microbial communities. These counts exhibit a sophisticated dependency structure inherited from the complicated interactions of the microbes. One source of domain knowledge in the microbiome context is the evolutionary relations among the microbes, summarized in a phylogenetic tree. Following this discussion, in Chapter 3 and Chapter 4 we develop methods for two classic statistical problems: the two-sample problem, and the clustering problem, in the microbiome context. Although numerous generic statistical tools have been proposed to solve these problems in various settings, their

performance is typically disappointing when directly applied to the microbiome amplicon sequencing data.

In Chapter 3, we introduce Bayesian Graphical Compositional Regression (BGCR) for testing the existence of and giving characterization to differences in the microbiome composition across groups of samples. Two important challenges of this problem are the large within-group heterogeneities among samples and the existence of potential confounding variables that, when ignored, increase the chance of false discoveries and reduce the power for identifying true differences. Our model utilizes a Pòlya tree type decomposition of the Dirichlet-multinomial distribution into a series of beta-binomial distributions along a phylogenetic tree, which provides a divide-and-conquer strategy that transforms the original high dimensional problem into multiple testing with a set of local tests at the internal nodes of the tree. To enhance the power of the local tests, we use a graphical model to achieve information sharing among them. Our framework adjusts for potential confounders in each local test and can be merged with the Bayesian testing strategy to give fast inference without the help of Monte Carlo simulations. The tree-based decomposition also provides a natural way to report and illustrate the testing result through reporting the nodes with a significant difference along the phylogenetic tree. With the proposed method, we analyze a version of the dataset from the American Gut Project (McDonald et al., 2018) to compare the gut microbiome composition of groups of participants with different dietary habits. Our analysis shows that (i) the frequency of consuming fruit, seafood, vegetable, and whole-grain is closely related to the gut microbiome composition and (ii) the conclusion of the analysis can change drastically when different sets of relevant covariates are adjusted, indicating the necessity of carefully selecting and including possible confounders in the analysis when comparing microbiome compositions with data from observational studies. This chapter is based on the published paper Mao et al. (2020).

In Chapter 4, we propose a model-based clustering method, referred to as Dirichlet-tree Multinomial Mixtures (DTMM), for microbiome data. DTMM can be viewed as a generalization to the Dirichlet Multinomial Mixtures (DMM)(Holmes et al., 2012), which clusters microbiome data with a finite mixtures of Dirichlet distributions. There are three key differences between DTMM and DMM. Firstly, DTMM induces a clustering pattern among samples with a Dirichlet process mixtures (Escobar and West, 1995), in which the number of clusters is not specified *a priori*. Secondly, DTMM utilizes the Dirichlet-tree distribution (Dennis, 1991) as the kernel for the mixtures. Unlike the Dirichlet distribution which has a single dispersion parameter to account for the within-cluster dispersions among samples, the number of dispersion parameters in a Dirichlet-tree distribution grows linearly with the number of categories in the microbiome data. Therefore, DTMM offers a better characterization of the "shapes" of the cluster. Thirdly, DTMM introduces a node-selection subroutine into the clustering procedure. Scientifically, it is important to identify the nodes that are actually contributing to the clustering. Statistically, this subroutine avoids unnecessary cluster-specific parameters and thus prevents the model from overfitting. We also analyze the American Gut Project data with DTMM. Specifically, we use DTMM to cluster two subsets of samples: those diagnosed with Inflammatory bowel disease and those with diabetes. Our analysis suggests that the identified clusters for different groups of samples can be different. Instead of determined by a few categories, the clusters of microbiome samples are typically determined by numerous categories jointly in a complicated manner. This chapter is based on the paper Mao and Ma (2020).

Determining the effect of a treatment is a central task in scientific research. In Chapter 5, we introduce a general framework for causal inference through Bayesian recursive partitioning, which induces a tree-structure. Popular methods for causal inference typically proceed with a prospective perspective by conditioning on a set

of pretreatment covariates and provide models for either the treatment assignment indicators or the outcome (or both). In this work, we take a retrospective perspective and instead model the covariates and the outcome jointly conditioning on the treatment assignment status. With this perspective, the classic problem of making causal comparisons through balancing the covariates of different treatment groups can be achieved through directly comparing the covariate distributions. At first glance, modeling the multivariate pretreatment covariate seems to be much harder than modeling the binary treatment status. We note that to achieve covariate balancing, the covariate model should focus on aspects of the covariate distributions that are relevant to balancing while treating the rest of the distributions as nuisance. To this end, we adopt a flexible nonparametric model for the covariates that focuses on the relations of the covariate distributions under different treatment groups while allowing the nuisance part of the distributions to be integrating out easily. With this joint modeling framework, we propose two strategies for estimating the causal effect. Our first strategy is a two-step approach that first finds regions in the covariate space within which the covariate distributions are balanced, and then estimates the treatment effect conditioning on the identified balancing blocks. This approach can be viewed as a generalization to the classic propensity score stratification (Rosenbaum and Rubin, 1984). The second strategy jointly models the covariates and the outcome, offering a way towards estimating the heterogeneous treatment effect. With this approach, the treatment effect is obtained from the posteriors of the parameters in the outcome model.

2

Microbiome and Microbiome Data

2.1 Microbiome and microbiome research

The microbiome is the collective genomes of all microbes that inhabit a specific environment, be it an animal, a plant or a tissue. In particular, the human microbiome is the genomes of the microbes that inhabit the human body and is considered as a counterpart or an extension to the human genome (Grice and Segre, 2012). The microbes in human body make up to 100 trillion cells collectively, encoding 100-fold more unique genes than the human genome (Qin et al., 2010), making the human microbiome a huge tank of information. There are several features of the human microbiome that make it popular in recent scientific research. Firstly, it has significant influences on various aspects of our physiology (Turnbaugh et al., 2009; Karlsson et al., 2013). For example, diseases such as obesity and Type 2 diabetes have been shown to be related to the gut microbiome (Turnbaugh et al., 2006; Qin et al., 2012). Secondly, unlike the relatively stable human genome, the human microbiome is highly variable. Although functionally similar, the microbe compositions of two individuals can be significantly different. Even for a single individual, the composi-

tion of the microbiome can be mutable spatially and temporally (Kuntz and Gilbert, 2017). Thirdly, the microbiota interacts with the body environment and can react to changes of the host sites. For example, alternations of the gut microbiome can be made by antibiotic administrations or dietary changes (Lozupone et al., 2012). For these reasons, learning and understanding the human microbiome provides a way to collect information about various aspects of human health and a way towards precision medicine (Kuntz and Gilbert, 2017).

The development of next-generation sequencing strategies enables us to profile the microbiome fast and economically, through either amplicon sequencing on target genes (usually the 16S ribosomal RNA gene) or shotgun sequencing on the entire microbial genome. Traditionally, the sequencing reads are sent to preprocessing pipelines such as QIIME (Caporaso et al., 2010) to construct clusters named operational taxonomic units (OTUs) based on certain predefined similarity threshold (typically 97%). In contrast, more recently developed pipelines such as DADA2 (Callahan et al., 2016) directly resolve amplicon sequence variants (ASVs), which is shown to outperform OTUs in terms of accuracy and interpretability (Callahan et al., 2017). OTUs and ASVs serve as the unit for the downstream statistical analyses of the microbiome and provide the same interface: each sample can be written a vector of counts on a list of units (OTUs or ASVs), representing the composition of the underlying community. The methodologies developed in the following chapters depend only on the interface and apply to both OTUs and ASVs, we thus use the customary "OTU" to refer to the unit.

2.2 Microbiome data and modeling

A preprocessed microbiome dataset that is ready for statistical analyses typically consists of three parts: a table of the OTU counts, a phylogenetic tree representing the evolutionary relationship of the OTUs, and a set of covariates for each sample.

Microbiome studies often collect covariate information about the host individual of each sample. For example, these covariates may include the demographic information of the participants, their dietary habits, and histories of disease. These information are sometimes collected via questionnaires on a voluntary basis. It is not uncommon for a microbiome study to be observational. Thus the distribution of the covariates are generally out of the control of the researchers.

The key part of a microbiome dataset is the OTU table, which summarizes the count of each OTU in each sequenced sample and provides a snapshot of the underlying microbial community. Microbiome count data has several features: (i) compositional: the OTU counts are compositional in nature. The total OTU count in a sample is artificial and can depend on various factors such as the sequencing depth. Therefore, the values of the OTU counts are meaningful only proportionally. (ii) high dimensional: the number of OTUs in a microbiome dataset is often large. A typical microbiome dataset contains tens of thousands of different categories. (iii) sparsity: the total OTU count of a sample is dominated by a few OTUs, with others having counts closed or equal to zero, referred to as the "rare biosphere" (Sogin et al., 2006). The zeros can either represent the absence of taxa from a particular sample or result from under-sampling (Tsilimigras and Fodor, 2016). (iv) complex covariance structure: the OTUs in a microbial community are functionally and evolutionarily related in a complex manner. As a result, the correlations among counts of different OTUs are also complicated. (v) large cross-sample variations: the OTU counts of samples often show large within-group heterogeneities.

These features of the microbiome count data poses challenges to the downstream statistical analyses, both computationally and from a modeling perspective. As we shall see in Chapter 3 and Chapter 4, off-the-shelf statistical tools developed for general purposes are typically not suitable for analyzing microbiome datasets directly. In this work, we focus on analyzing microbiome datasets with generative

models. A proper generative model that gives descent description of the microbiome count data must take into account these features.

There are two classic models for compositional data in general: the Dirichlet-multinomial (DM) model and the logistic-normal multinomial (LNM) model (Atchison and Shen, 1980). Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)$ be a vector of counts and suppose that $\boldsymbol{y} \sim \text{Multi}(N, \boldsymbol{p})$. DM and LNM are obtained by choosing different models for $\boldsymbol{p}$. Specifically, we have

- *DM*: $\boldsymbol{p} \sim \text{Dir}(\boldsymbol{a})$.
- *LNM*:

$$\boldsymbol{x} = \left(\log \frac{p_1}{p_M}, \log \frac{p_1}{p_M}, \ldots, \log \frac{p_{M-1}}{p_M} \right)$$

$$\boldsymbol{x} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Both DM and LNM have been applied to the microbiome context to answer various statistical queries (La Rosa et al., 2012a; Holmes et al., 2012; Xia et al., 2013; Silverman et al., 2019). When applied to the microbiome count data directly, DM suffers from the lack of flexibility in modeling the covariance structure among different categories, reflected by the single dispersion parameter of the Dirichlet distribution. The LNM, on the other hand, offers a very flexible treatment to the covariance structure. However, it entails a pre-selected base OTU that breaks the symmetry of the categories and doing inference under such a flexible setup is challenging. Moreover, direct application of DM and LNM to microbiome data ignores the biological relations of the OTUs. Incorporate such information requires extra tailors to these models, usually in an indirect manner. For example, Silverman et al. (2019) uses the phylogenetic distance among the OTUs to guide the prior selection of the covariate matrix for a LNM-type model.

In this work, we explore the so-called Dirichlet-tree multinomial (DTM) model

(Wang and Zhao, 2017; Tang et al., 2018), which is a tree-based extension to the DM model. Specifically, DTM utilizes a decomposition of the multinomial likelihood into a series of binomial likelihood along the tree and brings in a local beta-binomial model at each internal node (non-leaf node) of the tree. DTM balances the model flexibility and the computation feasibility—it generalizes DM to offer a richer covariance structure among the categories while not deviating "too far" from DM to keep inference practical. For a count vector with M categories, DM has only 1 parameter for the dispersions while LNM has $M(M-1)/2$. In comparison, DTM offers $(M-1)$ free dispersion parameters when the tree is a full binary tree. Clearly, the covariance structure offered by DTM is still restrictive compared to LNM especially when M is large, which is typical in microbiome studies. A heuristic illustration of the relations of the three models are shown in Figure 2.1. Although LNM is much more flexible than DTM in general, some DTM (even DM) cannot be well approximated by LNM (Hoff, 2003; Atchison and Shen, 1980)

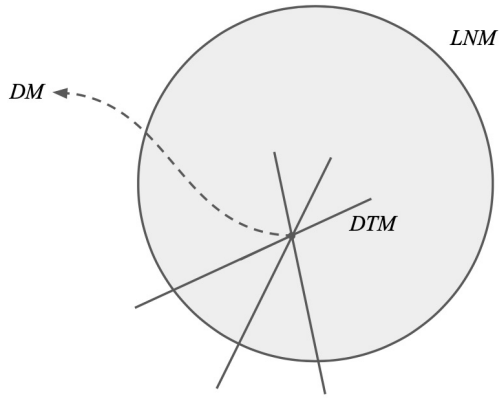

FIGURE 2.1: DM, LNM and DTM. This illustration is only heuristic. Each segment represent a family of DTMs for a specific tree structure.

To ensure that the additional $(M-2)$ dispersion parameters over DM are spent efficiently, the tree in DTM must be a "good" one—it has to provide useful infor-

mation about the correlations among the OTU counts. In the microbiome context, a natural candidate of such a tree is the phylogenetic tree, which summarizes the evolution hierarchy of the OTUs with a tree. As a simple illustration, we consider the top 20 OTUs based on total counts in a microbiome dataset (introduced in the next section). The phylogenetic tree over these OTUs are shown in Figure 2.2 (left). Figure 2.2 (right) shows two scale-free association measure—the Spearman's ρ and the Kendall's τ—between pairs of OTUs. Clearly, the marginal associations between pairs of OTU are partly reflected by the phylogenetic tree.

We also consider conditional associations between pairs of OTUs given other OTUs. Figure 2.3 (left) shows the estimated conditional dependency coefficients defined through a semi-parametric copula model as in (Hoff et al., 2007). Green edges represent positive associations, red edges represent negative associations. The darkness and the width of an edge shows the strength of the conditional association described by that edge. To better visualize the pattern of the conditional associations, we remove the edges in Figure 2.3 (left) with absolute value less than 0.1. The resulting plot is shown in Figure 2.3 (Right). Generally, counts of OTUs that are closer in the phylogenetic tree tend to have stronger correlations. The three separated clusters in Figure 2.3 (Right) are well represented by the three clusters obtained by cutting the phylogenetic tree at a high level (the dashed line in Figure 2.2 (left)). Note that unlike DM and LNM, biological information are incorporated directly into the generative modeling framework under DTM with the phylogenetic tree.

Computationally, DTM allows fast and efficient inference through a divide-and-conquer strategy. Specifically, the tree-decomposition transforms the original high-dimensional computational problem into a series of lower-dimensional ones, which can be solved in a relatively easy way and in parallel. This point would be recurrent in the following chapters.

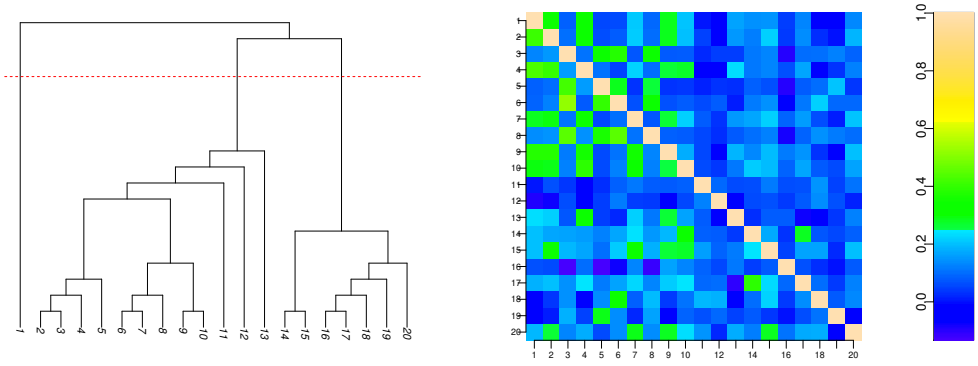

FIGURE 2.2: Left: the phylogenetic tree over the top 20 OTUs. Right: correlation measures of the OTU pairs.

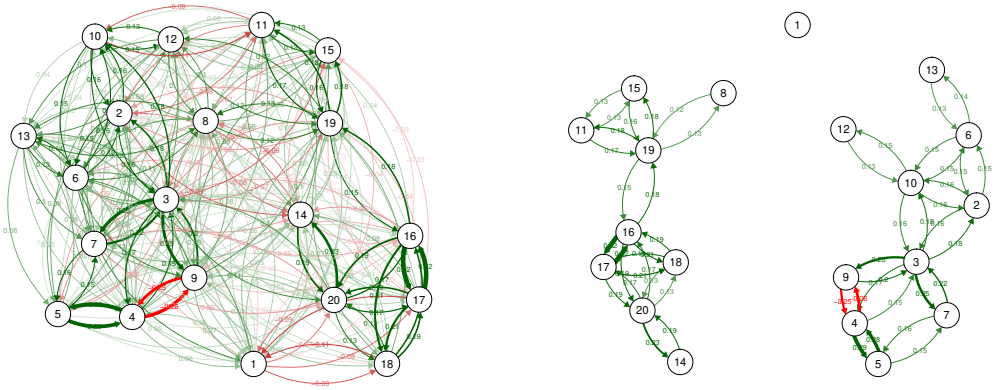

FIGURE 2.3: Left: Estimated conditional dependency coefficients among the OTUs (only coefficients with posterior 95% credible intervals do not cover zero are shown). Right: the estimated conditional dependency coefficients thresholded at 0.1.

2.3 The American Gut project

The American Gut (AG) project (McDonald et al., 2015a, 2018) aims at building an open-source and open-access reference microbiome dataset for general scientific use based on 16S rRNA sequencing and the QIIME pipeline (Caporaso et al., 2010). It collects mouth, skin and feces samples over a large variety of US participants on a voluntary basis. The participants send their microbiome samples to UC San Diego for sequencing and complete a questionnaire that covers the dietary habits, lifestyle, and health history. The AG project offers publicly accessible datasets that can be downloaded from `http://americangut.org`. In this work, we use the July 29, 2016 version of the fecal data of the AG project to illustrate the proposed methods. The full dataset we use contains counts of 27774 OTUs from 8327 fecal samples. Each sample is taken from a unique participant individual; 455 covariates describing various aspects of the host individuals are collected, such as their demographical information and dietary behaviors. Moreover, a rooted full binary phylogenetic tree on all OTUs is available.

3

Bayesian Graphical Compositional Regression

3.1 Introduction

An important task of microbiome studies is to compare the composition of the microbial community of groups of subjects (Hildebrandt et al., 2009; Wu et al., 2011; Qin et al., 2012; David et al., 2014). Various methods have been proposed for comparing two groups of data samples and their underlying distributions, ranging from the classic t-test to numerous recently developed methods (see for example Holmes et al. (2015), Soriano and Ma (2017) and Chen and Friedman (2017)). However, these generic approaches are either inapplicable or severely underpowered when the samples are OTU counts. In this chapter, we analyze the American Gut data (McDonald et al., 2015b) to compare the gut microbiome composition of groups of people with different dietary habits. The American Gut data includes samples from a large population and provides abundant sample-specific covariate information. Previous analyses of the American Gut data that compare groups of microbiome samples can be improved since they either focus on certain summary statistics of the microbiome compositions (Goedert et al., 2014), or disregard the observational nature of the

American Gut project by ignoring possible confounders (Tang et al., 2018). Our approach aims at addressing these limitations.

As introduced in Chapter 2, the specific features of the OTU counts make this two-sample-problem challenging in the microbiome context. In this chapter, we propose a testing method, called Bayesian graphical compositional regression (BGCR), that utilizes both the phylogenetic information among the OTUs and the covariate information of the host participants to tackle the aforementioned modeling and computational challenges.

BGCR aims to effectively account for the specific features of microbiome data by combining three techniques—(i) a decomposition of the multinomial likelihood along the phylogenetic tree, (ii) graphical models , and (iii) the Bayesian testing framework. Broadly speaking, our approach falls into a stream of works based on the Dirichlet-multinomial model (DM) (La Rosa et al., 2012a) that puts a Dirichlet prior on the multinomial parameters to resolve the large overdispersions in OTU counts. Recent developments along the line use the phylogenetic tree—which summarizes the evolutionary relationship among the OTUs and thus serves as a proxy to their functional relationship—to enrich the model construction. Specifically, Wang and Zhao (2017) adopt the Dirichlet tree multinomial model (DTM) in which OTU counts are aggregated along the phylogenetic tree and a DM model is introduced to characterize how the OTU counts on each internal node of the tree are distributed into its child nodes. BGCR adopts a similar decomposition to transform the original testing problem into multiple testing involving a collection of node-specific tests along the phylogenetic tree.

Instead of treating the node-specific tests independently, we use a bottom-up graphical structure to introduce dependency among them. Our motivation is that some of these tests may have poor statistical power due to the limited number of samples and relevant OTU counts. The dependency structure allows effective in-

formation sharing among the node-specific problems, thereby improving the power of the tests. Our choice of a bottom-up autoregressive specification comes from the observation that compositional differences are often observed to cluster into "chains" along the phylogenetic tree (Tang et al., 2018). Such chains can arise in two ways: either due to a structural constraint of DTM—a cross-group difference at any particular node in the phylogenetic tree induces (weaker) differences in its ancestors, or as a result of the functional relatedness of OTUs in the phylogenetic tree.

The decomposition scheme and the graphical structure are merged with the Bayesian testing framework to give fast and interpretable inference through a suite of computational techniques including numerical integration and message passing. On the one hand, exact posterior summaries are available through a recursive algorithm that does not require Monte Carlo simulations; on the other, the testing results are given a full probabilistic characterization with the uncertainty quantified in a coherent manner. In addition, we show that the Bayesian test results in substantially improved power, likely due to the way the within-group variability is dealt with—integrated out rather than estimated as in existing approaches.

Besides the phylogenetic tree, covariate information is also used to assist the modeling of the microbome composition in recent works focusing on prediction (Tang and Nicolae, 2017) and selection of the covariates with significant associations to the OTU counts (Xia et al., 2013; Wang and Zhao, 2017; Wadsworth et al., 2017; Ren et al., 2017; Grantham et al., 2017). In contrast, little attention has been paid to the cross-group comparison scenario, where incorporating covariates is even more crucial since many microbiome studies are observational and the effects of unadjusted confounders could invalidate the testing results by creating false positives. As a starting point, BGCR allows a node-specific regression adjustment for important covariates. This adjustment can be incorporated in the inference mechanism with little extra computational burden and is able to reduce the number of false positives

effectively, as we will illustrate in our numerical examples. In our analysis of the American Gut data, we show that when a set of possible confounders are ignored, the testing results can be misleading.

In summary, in comparison to existing approaches, the contribution of BGCR are three-fold: (i) it allows a Bayesian framework for testing cross-group differences in microbiome composition along a phylogenetic structure; (ii) it provides a principled probabilistic modeling framework for borrowing information across taxa, thereby enhancing the power for detecting cross-group differences; and (iii) it uses a principled way to incorporate additional covariates in the Bayesian testing setting while maintaining the computational efficiency of the proposed framework.

To close the introduction, we connect BGCR to some relevant references in the DM literature stream. Using DM, La Rosa et al. (2012a) proposed a generalized Wald-type test statistic based on method-of-moments estimates of the Dirichlet parameters. Due to the large number of OTUs, this test has large degrees of freedom and is usually underpowered. One simple attempt to alleviate this problem is to aggregate OTUs from the same genus and study their composition at the genus level as suggested by Chen and Li (2013). However, one could also aggregate the OTUs to other levels in the hierarchy of biological classification, such as the family or order level, yet different levels of aggregation can result in inconsistent testing results (Tang et al., 2018). Similar to BGCR, Tang et al. (2018) proposed the PhyloScan test, which also decomposes the original problem to a series of node-specific tests along the phylogenetic tree and tries to incorporate dependencies among these tests. BGCR and PhyloScan differ in the way they deal with these underlying dependencies. PhyloScan introduces dependencies through a scanning procedure by adding up the test statistics in triplets of neighboring nodes while BGCR considers a probabilistic graphical model that describes the dependencies in a generative manner. The latter allows more flexible borrowing of information beyond neighboring triplets

along with a fully probabilistic interpretation.

In Section 3.2, we describe our method for testing and characterizing the cross-group differences of OTU compositions. In Section 3.3, four representative simulation scenarios are considered to evaluate the performance of the proposed method. An application of the method to the American Gut data is shown in Section A. Section 3.5 concludes.

3.2 Method

3.2.1 Data and background

In this section, we set up some notations and briefly review DM and DTM for OTU compositions. In this work, all three parts of the microbiome dataset are utilized.

OTU counts. Consider a microbiome dataset with OTU counts of two groups of subjects on M OTUs denoted by $\Omega = \{\text{OTU}_1, \text{OTU}_2, \ldots, \text{OTU}_M\} = \{\omega_1, \omega_2, \ldots, \omega_M\}$. Let n_i be the number of samples in group i, $i = 0, 1$. For the j-th sample in group i, $j = 1, \ldots, n_i$, let $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijM})$ be the vector of its OTU counts, where y_{ijl} is the number of the l-th OTU in this sample for $l = 1, \ldots, M$. Moreover, let $N_{ij} = \sum_{l=1}^{K} y_{ijl}$ be the total number of OTU counts in that sample. For simplicity, let $\boldsymbol{Y}$ be the OTU counts of all the samples. Table 3.1 illustrates the OTU counts for a certain group of samples.

Phylogenetic tree. Let $\mathcal{T}$ be a rooted full binary phylogenetic tree that describes the evolutionary relations of the M OTUs. Let $\mathcal{I}$ denote the set of its internal nodes (i.e., non-leaf nodes). We denote each node A of $\mathcal{T}$ by the set of its descendant OTUs. For example, $A = \{\omega_l\}$, $1 \leqslant l \leqslant K$ represents a leaf of $\mathcal{T}$ that contains a single OTU ω_i; $A = \Omega$ is the root containing all the OTUs. Since the tree is full, each internal node has exactly two children. For $A \in \mathcal{I}$, let A_l, A_r be the left and right children of A. If $A \neq \Omega$, let A_p be its parent and A_s be its sibling (i.e., the node in $\mathcal{T}$ that has the same parent as A). If A is a leaf, for the j-th sample in group i,

we let $y_{ij}(A) = y_{ijl}$ be the count of the OTU that A represents; if $A \in \mathcal{I}$, we recursively define the count in A to be the aggregation of the counts in its two children: $y_{ij}(A) = y_{ij}(A_l) + y_{ij}(A_r)$. Equivalently, $y_{ij}(A) = \sum_{\{l:\omega_l \in A\}} y_{ijl}$. Figure 3.1 shows an example of a phylogenetic tree over 6 OTUs. In this example, $A = \{\omega_3, \omega_4, \omega_5, \omega_6\}$, $y_{ij}(A) = y_{ij3} + y_{ij4} + y_{ij5} + y_{ij6}$.

Table 3.1: An $n_i \times M$ OTU table for group i.

Sample	ω_1	ω_2	$\cdots$	ω_M	Sum
1	y_{i11}	y_{i12}	$\cdots$	y_{i1M}	N_{i1}
2	y_{i21}	y_{i22}	$\cdots$	y_{i2M}	N_{i2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n_i	y_{in_i1}	y_{in_i2}	$\cdots$	y_{in_iM}	N_{in_i}

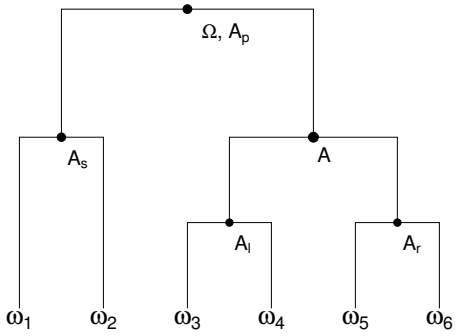

FIGURE 3.1: A phylogenetic tree over 6 OTUs.

Covariates. For the j-th sample in group i, let $\boldsymbol{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})' \in \mathbb{R}^{p+1}$ denote the p covariates that we want to include in the analysis. Moreover, let $z_{ij} \in \{0, 1\}$ be its group indicator.

La Rosa et al. (2012a) used the Dirichlet-multinomial model (DM) to account for the within-group heterogeneity among the OTU counts:

$$\begin{aligned} \boldsymbol{y}_{ij} \mid N_{ij}, \boldsymbol{\pi}_{ij} &\overset{\text{ind}}{\sim} \text{Multinomial}(N_{ij}, \boldsymbol{\pi}_{ij}) \\ \boldsymbol{\pi}_{ij} \mid \boldsymbol{\pi}_i, \nu &\overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\nu \boldsymbol{\pi}_i) \end{aligned} \tag{3.1}$$

for $i = 0, 1$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ with $\sum_{l=1}^{K} \pi_{il} = 1$, and $\nu > 0$ a dispersion parameter that controls the within-group variability. A limitation of DM is its induced correlations among OTUs—that they are independent up to the summation constraint (Wang and Zhao, 2017), which is unrealistic since the OTUs have inherent

and complicated relationships with each other, partly reflected in their evolutionary history summarized in a phylogenetic tree.

Wang and Zhao (2017) propose to adopt the Dirichlet-tree multinomial model (DTM) by directly incorporating the phylogenetic tree in the modeling of the OTU counts. Specifically, DTM models how the counts in $A \in \mathcal{I}$ are distributed to its two children by a beta-binomial model:

$$\begin{aligned} y_{ij}(A_l) \mid y_{ij}(A), \theta_{ij}(A) &\overset{\text{ind}}{\sim} \text{Binomial}(y_{ij}(A), \theta_{ij}(A)) \\ \theta_{ij}(A) \mid \theta_i(A), \nu(A) &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\theta_i(A)\nu(A), (1-\theta_i(A))\nu(A)), \end{aligned} \tag{3.2}$$

where $\theta_{ij}(A)$ denotes the proportion of counts in node A that are distributed to its left child A_l, $\theta_i(A)$ the mean of the $\theta_{ij}(A)$'s in group i, and $\nu(A)$ a precision or dispersion parameter that controls the variability of $\theta_{ij}(A)$ around the group mean, for $i = 1, 2$, $j = 1, \ldots, n_i$. Let $\mathcal{L}_{BB}(\theta_i(A), \nu(A) \mid y_{ij}(A_l), y_{ij}(A_r))$ denote the likelihood of the parameters in (3.2) obtained by integrating out $\theta_{ij}(A)$, which with slightly more general notations can be written as

$$\mathcal{L}_{BB}(\theta, \nu \mid y_1, y_2) = \begin{cases} B(\theta\nu + y_1, (1-\theta)\nu + y_2)/B(\theta\nu, (1-\theta)\nu), & \text{if } \nu < \infty \\ \theta^{y_1}(1-\theta)^{y_2}, & \text{if } \nu = \infty \end{cases} \tag{3.3}$$

for $y_1, y_2 \in \{0, 1, 2, \ldots, \}, \theta \in [0, 1]$ and $\nu \in (0, \infty]$, where $B(\cdot, \cdot)$ is the Beta function. It can be shown that the DM likelihood for each sample can be factorized into a series of beta-binomial likelihoods along the tree

$$\mathcal{L}_{DM}(\boldsymbol{\pi}_i, \nu \mid \boldsymbol{y}_{ij}) = \prod_{A \in \mathcal{I}} \mathcal{L}_{BB}(\theta_i(A), \nu_i(A) \mid y_{ij}(A_l), y_{ij}(A_r)) \tag{3.4}$$

provided that $\theta_i(A) = \pi_i(A_l)/\pi_i(A)$ and $\nu_i(A) = \nu\pi_i(A)$, where $\pi_i(A) = \sum_{l \in A} \pi_{il}$ for $A \in \mathcal{I}$ (Dennis, 1991, 1996). In this sense, DTM could be seen as a generalization of DM.

3.2.2 Bayesian compositional regression for microbiome data

The likelihood factorization (3.4) suggests a "divide-and-conquer" strategy to perform inference on $\boldsymbol{\pi}_i$. Specifically, inference on $\boldsymbol{\pi}_i$ could be achieved equivalently by doing inference on $\{\theta_i(A) : A \in \mathcal{I}\}$, $i = 0, 1$. In this section, we use this strategy to transform the original problem of comparing microbial compositions of two groups of sample into a series of tests on $\{\theta_i(A) : A \in \mathcal{I}\}$ in a more general situation that allows adjustment for covariates.

For each $A \in \mathcal{T}$, we modify the beta-binomial specification in DTM to allow regression adjustments for covariates:

$$\begin{aligned} y_{ij}(A_l) \mid y_{ij}(A), \theta_{ij}(A) &\overset{\text{ind}}{\sim} \text{Binomial}(y_{ij}(A), \theta_{ij}(A)) \\ \theta_{ij}(A) \mid \boldsymbol{x}_{ij}, z_{ij}, \nu(A) &\sim \text{Beta}(\theta_{\boldsymbol{x}_{ij}, z_{ij}}(A)\nu(A), (1 - \theta_{\boldsymbol{x}_{ij}, z_{ij}}(A))\nu(A)) \\ g(\theta_{\boldsymbol{x}_{ij}, z_{ij}}(A)) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}(A) + z_{ij}\gamma(A), \end{aligned} \tag{3.5}$$

where $g : [0, 1] \to \mathbb{R}$ is a link function, such as the logit link $g(x) = \log(x/(1 - x))$ for $x \in (0, 1)$. $\boldsymbol{\beta}(A) \in \mathbb{R}^{p+1}$ and $\gamma(A) \in \mathbb{R}$ are the unknown parameters of the local model on A. We refer to this model as the Bayesian compositional regression (BCR). Under BCR, testing the original null that there is no cross-group difference in the microbial composition is equivalent to jointly testing a set of local hypotheses on all $A \in \mathcal{I}$:

$$H_0(A) : \gamma(A) = 0 \quad \text{vs} \quad H_1(A) : \gamma(A) \neq 0. \tag{3.6}$$

Instead of modeling the OTU counts around some "group-specific centroid" as in DM and DTM, we model them around some "covariate-specific centroid" with the grouping information included as a special "covariate". Recent literature has shown that many covariates have associations and possible effects on the microbiome composition (Wang and Zhao, 2017; Wadsworth et al., 2017; Xia et al., 2013; Tang and Nicolae, 2017; Ren et al., 2017; Grantham et al., 2017). In the testing scenario,

some of these covariates may also be correlated with the grouping variable. Ignoring such confounding can reduce the power for identifying differences, or worse yet, lead to false positives. The regression adjustments in (3.5) block out the possible confounders and give us an opportunity to improve the accuracy of the testing results and reduce the chance of false discoveries.

We shall take a Bayesian approach to the local testing problem (3.6) for three reasons: (1) it provides a natural way to deal with the dispersion $\nu(A)$, essentially a nuisance parameter, through integration instead of potentially unstable points estimates; (2) it allows the introduction of dependencies among the local hypotheses in a generative manner as shown in Section 3.2.3; and (3) it gives a fully probabilistic characterization of the uncertainty.

To this end, we first need to specify priors for $\boldsymbol{\beta}(A)$ and for $\gamma(A)$ under $H_1(A)$. A simple choice is to put independent normal priors on the elements of $\boldsymbol{\beta}(A)$ and $\gamma(A)$. Note that using vague proper priors on the model specific parameters $\gamma(A)$ is often problematic in testing and would cause the so-called "Bartlett's Paradox" (Berger et al., 2001). Alternatively, various principles for constructing "objective" priors for the coefficients can be adopted. For example, generalizations of the g-prior and the mixture of g-priors (Liang et al., 2008) for linear regressions to GLMs can be employed here, see for example Held et al. (2015) and Li and Clyde (2015). Specifically, one could apply the local information metric (LIM) g-prior as suggested by Li and Clyde (2015) on $\gamma(A)$. In our setting, the difference in dimensions between the parameter space under the alternative and the null is only one; therefore, putting normal prior $\mathrm{N}(0, \sigma^2_\gamma(A))$ with reasonably large variance (such as $\sigma^2_\gamma(A) = 10$) would give reasonable results. For the "common" parameters $\boldsymbol{\beta}(A)$, we adopt independent normal priors $\mathrm{N}(0, \sigma^2_{\boldsymbol{\beta}}(A))$ with large $\sigma_{\boldsymbol{\beta}}(A)$ on its elements. For example, when the covariates are standardized (rescaled to have a mean of zero and a standard deviation of one), $\mathrm{N}(0, 16)$ covers most probable values of $\boldsymbol{\beta}(A)$ in many applications.

Without further knowledge about the nuisance dispersion parameter $\nu(A)$, we put a prior $G_A(\nu)$ on it and integrate it out. This is key to the improvement gained by applying a Bayesian testing scheme compared to its frequentist counterparts that rely on a point estimator of $\nu(A)$ (4.3). It is necessary for $G_A(\nu)$ to have a large support to allow various levels of dispersions. We hence take $\log_{10}\nu(A) \sim \text{Unif}(-1, 4)$ that covers a wide range of dispersion levels but does not put too much prior mass on unreasonably large or small values (Ma and Soriano, 2018). Other choices with unbounded support such as Gamma priors on $\nu(A)$ can also be adopted.

To perform posterior inference, we introduce the state indicator of the test on node A for $A \in \mathcal{I}$:

$$S(A) = \begin{cases} 0, & H_0(A) \text{ is true,} \\ 1, & H_1(A) \text{ is true.} \end{cases} \tag{3.7}$$

We assume for now that the $S(A)$'s are independent and let $\Pr(S(A) = 1) = \rho(A)$ *a priori*. The testing problem on A now becomes an inference problem on $S(A)$. Let $M_s(A)$ be the marginal likelihood under $H_s(A)$, $s \in \{0, 1\}$. We have

$$M_s(A) = C \iiint \prod_{i=1}^{2} \prod_{j=1}^{n_i} \mathcal{L}_{BB}(g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + z_{ij}\gamma), \nu \mid y_{ij}(A_l), y_{ij}(A_r))$$

$$dF_{s,A}(\boldsymbol{\beta}) dF_{s,A}(\gamma) dG_A(\nu),$$

where C is a constant with respect to the parameters, $F_{s,A}(\boldsymbol{\beta}), F_{s,A}(\gamma)$ are the priors for $\boldsymbol{\beta}(A)$ and $\gamma(A)$ under $H_s(A)$, $\mathcal{L}_{BB}(\cdot)$ the beta-binomial marginal likelihood defined in (3.3), and $G_A(\nu)$ the prior on $\nu(A)$. We give details on the computational strategy for evaluating $M_s(A)$ in Appendix A.1. Given $M_s(A)$, the posterior of $S(A)$ is

$$S(A) \mid \tilde{\rho}(A) \sim \text{Bernoulli}(\tilde{\rho}(A))$$

$$\tilde{\rho}(A) = \frac{\rho(A) M_1(A)}{(1 - \rho(A)) M_0(A) + \rho(A) M_1(A)}.$$

We shall refer to $\tilde{\rho}(A) = \Pr(S(A) = 1 \mid \boldsymbol{Y})$ as the **posterior marginal alternative probability** (PMAP) on node A, denoted also as PMAP(A). The PMAP is large when the evidence against the local null is strong. The set of PMAPs along $\mathcal{T}$ can be used to test the original null and pinpoint the differences. We postpone the details on using PMAPs for making such decisions based on multiple testing considerations to Section 3.2.5.

We end this section with three further comments on our model specification in BCR. Firstly, we assume that the two groups share a common dispersion parameter $\nu(A)$ to simplify the computation. If this assumption is not likely to hold in the data, group-specific dispersions can be adopted. Secondly, BCR does not automatically select covariates. This is not restrictive in the testing scenario since our primary goal is not to identify the best predictive model for the microbiome composition or to estimate the "effect size" of specific covariates on the microbiome composition, but to preclude them from introducing bias to the analysis. In practice, all the suspect confounders can be included in the model even some of them may turn out to have no confounding effects provided that the number of covariates is not too large. We note that one must practice caution in adopting covariate selection in this context. Traditional statistical variable selection is based on the predictive ability of the covariates, and can lead to inappropriate elimination of confounders and result in false positives. We provide more discussion and a simple numerical illustration in Appendix A.3. Thirdly, in the regression setup, we assume that the regression coefficients $\boldsymbol{\beta}(A)$ are the same for both groups. This "common slope" assumption essentially serves as an identifiability constraint, without which the null hypothesis is ill-defined—the two groups will be identical only for a specific combination of covariate values but not others, and consequently it would not make sense to test whether two groups of samples involving different covariate values are "identical" in microbiome composition.

3.2.3 Bayesian graphical compositional regression

In BCR, the test on node A is performed based only on the empirical evidence in A. Let $\mathcal{S} = \{S(A) : A \in \mathcal{I}\}$, the collection of state indicators of all internal nodes in $\mathcal{T}$. Elements in $\mathcal{S}$ are independent *a priori*. However, this independence assumption disregards the inherent relations of the $S(A)$'s—that they are naturally linked by the phylogenetic tree. Introducing suitable dependency structures on $\mathcal{S}$ can enhance inference because (1) cross-group differences often occur along the phylogenetic tree in a clustered manner, forming chains, which is pointed out in Tang et al. (2018) and will be confirmed in our data analysis, and (2) from a statistical perspective, some local tests may involve only few counts and their power is thus limited, especially when the nodes are close to the leaves of $\mathcal{T}$. Introducing a dependency structure on $\mathcal{S}$ allows these nodes to borrow information from each other and thus increases the power of the tests. The dependency structure we introduce should satisfy two desiderata. On the one hand, it should allow flexible information sharing among nodes that are close on the phylogenetic tree. On the other hand, it should be simple enough to keep the posterior inference tractable.

Recall that the node-specific beta-binomial models in BCR are naturally linked together by the phylogenetic tree, which provides a proxy to their functional relationship. This suggests that we can use the phylogenetic information when imposing dependency structures on $\mathcal{S}$. In particular, we consider a "bottom-up" auto-regressive structure such that if the signal (cross-group difference) is present at a certain node $A \in \mathcal{T}$, there is a chance for it to be "carried upwards" to A_p. Formally, for $A \in \mathcal{I}$, we let

$$\begin{aligned}\operatorname{logit}[\Pr(S(A) = 1 \mid S(A_l), S(A_r))] = \alpha(A) &+ \tau(A) \cdot \mathbb{1}_{[S(A_l)+S(A_r)\geqslant 1]} \\ &+ \kappa(A) \cdot \mathbb{1}_{[S(A_l)+S(A_r)=2]},\end{aligned} \tag{3.8}$$

where $\tau(A), \kappa(A) \geqslant 0$ together describe how likely signals at A_l or A_r are passed

upwards along the tree. If A_l or A_r (or both) is a leaf node, we set $S(A_l)$ or $S(A_r)$ to be zero since there is no test performed on that node. This model can be embedded in a family of models specified by the conditional probabilities

$$\Pr(S(A) = s \mid S(A_l) = s_l, S(A_r) = s_r) = \rho_{s_l s_r, s}(A), \tag{3.9}$$

where $s, s_l, s_r \in \{0, 1\}$ and $\rho_{s_l s_r, s} \in [0, 1]$. Specifically, the auto-regressive model forces $\rho_{01,s}(A) = \rho_{10,s}(A)$ since usually there is no specific prior information to differentiate between the two children of a node. Equivalently, the conditional probability of $S(A)$ given $S(A_l), S(A_r)$ can be characterized by a transition matrix

$$\boldsymbol{\rho}(A) = \begin{pmatrix} \rho_{00,0}(A) & \rho_{00,1}(A) \\ \rho_{01,0}(A) & \rho_{01,1}(A) \\ \rho_{10,0}(A) & \rho_{10,1}(A) \\ \rho_{11,0}(A) & \rho_{11,1}(A) \end{pmatrix}$$

where the elements in each row of $\boldsymbol{\rho}(A)$ sum up to 1. We denote $\boldsymbol{\rho} = \{\boldsymbol{\rho}(A) : A \in \mathcal{I}\}$ and defer the specification of elements in $\boldsymbol{\rho}$ to Section 3.2.5. The set of auto-regressive models with transition probabilities $\boldsymbol{\rho}$ induces a joint distribution $F_{\mathcal{S}}$ on $\mathcal{S}$. Incorporating this distribution into BCR leads to a hierarchical formulation of our model for each $A \in \mathcal{I}$:

$$\begin{aligned}
S(A) \mid \boldsymbol{\rho} &\sim F_{\mathcal{S}}, \quad \nu(A) \mid G_A \overset{\text{ind}}{\sim} G_A \\
\boldsymbol{\beta}(A) &\overset{\text{ind}}{\sim} \mathrm{N}_{p+1}(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}(A) \cdot I_{p+1}) \\
\gamma(A) &\overset{\text{ind}}{\sim} \mathbb{1}_{[S(A)=0]} \cdot \delta_0 + \mathbb{1}_{[S(A)=1]} \cdot \mathrm{N}(0, \sigma^2_{\gamma}(A)) \\
g(\theta_{\boldsymbol{x}_{ij}, z_{ij}}(A)) &= \boldsymbol{x}_{ij}^{\top} \boldsymbol{\beta}(A) + z_{ij}\gamma(A) \\
\theta_{ij}(A) \mid \boldsymbol{x}_{ij}, z_{ij}, \nu(A) &\sim \mathrm{Beta}(\theta_{\boldsymbol{x}_{ij}, z_{ij}}(A)\nu(A), (1 - \theta_{\boldsymbol{x}_{ij}, z_{ij}}(A))\nu(A)) \\
y_{ij}(A_l) \mid y_{ij}(A), \theta_{ij}(A) &\overset{\text{ind}}{\sim} \mathrm{Binomial}(y_{ij}(A), \theta_{ij}(A)), \quad i = 1, 2, j = 1, \ldots, n_i,
\end{aligned} \tag{3.10}$$

where I_{p+1} is the $(p+1)$-dimensional identity matrix, $\mathbb{1}_{[\cdot]}$ the indicator function,

and δ_0 a point mass at zero. We shall refer to this model as the Bayesian graphical compositional regression (BGCR).

An alternative way to introduce dependencies among elements in $\mathcal{S}$ that also incorporate the phylogenetic information is by utilizing the "top-town" Markov tree model (MT) (Crouse et al., 1998; Soriano and Ma, 2017), which instead specifies $\Pr(S(A) \mid S(A_p))$ for $A \in \mathcal{T}\backslash\{\Omega\}$. However, the "explaining away" effect (Wellman and Henrion, 1993) of the auto-regressive model is important for microbiome data. MT pushes signals downwards along the tree to both children of a node and implies cooccurrence of signals in the sibling nodes, which is not the typical situation in the microbiome context. Instead, the signals in microbiome studies often form chains, with only one of the two children nodes share the signal with the parent node. This is the primary reason for adopting a "bottom-up" autoregressive model.

3.2.4 Inference under BGCR

As in BCR, the marginal posteriors $\{\Pr(S(A) = 1 \mid \boldsymbol{Y}) : A \in \mathcal{I}\}$ play a pivotal role in the posterior inference under BGCR. We next show that these quantities can be calculated exactly up to the approximation to $M_s(A)$'s, without entailing Monte Carlo simulations.

The joint distribution introduced by BGCR on $\mathcal{S}$ can be represented by a Bayesian network, i.e., a directed acyclic graph (DAG) $\mathcal{G}(V_{\mathcal{S}}, B_{\mathcal{S}})$ with nodes $V_{\mathcal{S}} = \{S(A) : A \in \mathcal{I}\}$ and edges $B_{\mathcal{S}} = \{S(A) \to S(A_p) : A \in \mathcal{I}\backslash\{\Omega\}\}$. Moralizing $\mathcal{G}(V_{\mathcal{S}}, B_{\mathcal{S}})$ yields an undirected graph $\mathcal{U}(V_{\mathcal{S}}, E_{\mathcal{S}})$ from which the conditional dependencies in $\mathcal{S}$ can be read off directly (Koller and Friedman, 2009). Specifically, the vertices of $\mathcal{U}$ are the same of those of $\mathcal{G}$; there is an undirected edge between $S(A)$ and $S(A')$ in $\mathcal{U}$ if A and $A' \in \mathcal{I}$ are siblings or if they form a parent-child pair: $E_{\mathcal{S}} = \{\{S(A), S(A')\} : A' = A_p \text{ or } A' = A_s, A \in \mathcal{I}\backslash\{\Omega\}\}$. Moreover, we refer to a chain of edges as a "path". Two nodes $S(A)$ and $S(A')$ in $\mathcal{U}$ are conditionally independent

given a set of nodes $v = \{S(A_1), \ldots, S(A_m)\} \subset V_{\mathcal{S}}$ if and only if any path connecting $S(A)$ and $S(A')$ passes some elements in v. Figure 3.2 (i), (ii), and (iii) respectively give illustrations of a fictional phylogenetic tree, the corresponding Bayesian network $\mathcal{G}$ on $\mathcal{S}$, and the moralized undirected graph $\mathcal{U}$. The structure of $\mathcal{U}$ induced by BGCR has two implications on posterior computations. On the one hand, $\mathcal{U}$ contains many loops, which at first glance would substantially complicate the posterior and the inference algorithm. If $\mathcal{U}$ were a tree-type graph without loops, as in MT, simple and exact inference can be achieved by a forward-backward algorithm. On the other hand, all the loops in $\mathcal{U}$ include only three nodes and the dependencies among $S(A)$'s are local—$S(A)$ is independent of other elements in $\mathcal{S}$ once conditioning on $S(A_p), S(A_s), S(A_l)$ and $S(A_r)$, suggesting that $\mathcal{U}$ is not "too far" from a tree-type graph that allows exact inference. Intuitively, if we treat each loop as a unit, $\mathcal{U}$ can be viewed as a "tree of loops" on which exact inference can now be performed.

Formally, for $S(A) \in V_{\mathcal{S}}$, we call $C(A) = (S(A), S(A_l), S(A_r))$ the **clique** of $S(A)$, or simply the **clique** of A. The conditional dependencies among $\mathcal{C} = \{C(A) : A \in \mathcal{I}\}$ are inherited from $\mathcal{U}$ and can be represented by an undirected graph $\mathcal{J}(V_{\mathcal{C}}, E_{\mathcal{C}})$ in a similar manner, where $V_{\mathcal{C}} = \mathcal{C}$, $E_{\mathcal{C}} = \{\{C(A), C(A')\} : A' = A_p, A \in \mathcal{I} \backslash \{\Omega\}\}$. See Figure 3.2 (iv) for an illustration. It is easy to see that $\mathcal{J}$ contains no loop and that any two nodes in $\mathcal{J}$ are connected by a unique path. Therefore, $\mathcal{J}$ is an undirected tree. We refer to $\mathcal{J}$ as the **clique tree** and perform inference thereon. Note that $C(A)$ can take on eight values, denoted by $\boldsymbol{c} = \{c_1, c_2, \ldots, c_8\}$ where $c_i = (c_{i1}, c_{i2}, c_{i3}) = (\lfloor \frac{i-1}{4} \rfloor \bmod 2, \lfloor \frac{i-1}{2} \rfloor \bmod 2, (i-1) \bmod 2)$ (each element is the corresponding digit of the binary representation of i).

For $A \in \mathcal{I}$, without loss of generality, let A be the left child of A_p. Eq. (3.9)

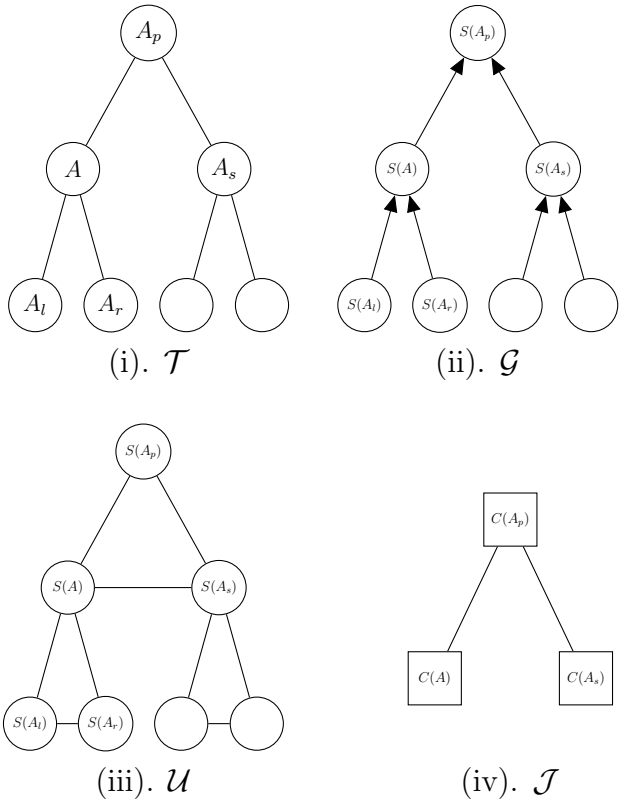

FIGURE 3.2: (i). A fictional phylogenetic tree (the leaves of $\mathcal{T}$ are not shown); (ii). The "bottom-up" Bayesian network $\mathcal{G}$ on $\mathcal{S}$ corresponds to $\mathcal{T}$; (iii). The undirected graph after moralizing $\mathcal{G}$; (iv). The clique tree derived from $\mathcal{U}$.

induces the top-down transition probabilities on the clique tree $\mathcal{J}$:

$$\begin{aligned}\Pr(C(A) = c_{i'} \mid C(A_p) = c_i) &= \Pr((S(A), S(A_l), S(A_r)) = c_{i'} \mid (S(A_p), S(A), S(A_s)) = c_i) \\ &= \Pr((S(A_l), S(A_r)) = (c_{i'2}, c_{i'3}) \mid S(A) = c_i),\end{aligned} \tag{3.11}$$

for $C(A) \in \mathcal{C}$, where $i, i' = 1, 2, \ldots, 8$. As with $\boldsymbol{\rho}(A)$, these transition probabilities could be organized into an 8×8 matrix $\boldsymbol{\xi}(A)$ with the i'-th element in row i defined by $\boldsymbol{\xi}_{ii'}(A) = \Pr(C(A) = c_{i'} \mid C(A_p) = c_i)$. Each row of this matrix represents the conditional distribution of $C(A)$ when $C(A_p)$ takes one of the eight possible values in $\boldsymbol{c}$. Since $\Omega \in \mathcal{T}$ does not have a parent, we let each row of $\boldsymbol{\xi}(\Omega)$ be the induced marginals of $C(\Omega)$.

In addition, for $A \in \mathcal{I}$, let $\mathcal{T}(A)$ be the subtree of $\mathcal{T}$ with A as the root. Recall that for the j-th sample in group i, $y_{ij}(A)$ denotes the OTU counts in node A. Let $\boldsymbol{y}_{ij}(A)$ denote all the OTU counts that fall into the subtree $\mathcal{T}(A)$. Note that the difference between $y_{ij}(A)$ and $\boldsymbol{y}_{ij}(A)$ is important: $y_{ij}(A)$ contains only the counts in A while $\boldsymbol{y}_{ij}(A)$ contains the set of $y_{ij}(A_d)$'s for all $A_d \in \mathcal{T}(\mathcal{A})$. Let $\boldsymbol{Y}(A) = \{\boldsymbol{y}_{ij}(A) : i = 1, 2, j = 1, \ldots, n_i\}$ be the counts in $\mathcal{T}(A)$ from all the samples. Bayesian inference on $S(A)$ relies on evaluating the posterior transition probability matrix $\tilde{\boldsymbol{\xi}}(A)$ for each $C(A) \in \mathcal{C}$. We next describe how these posteriors can be calculated with a recursive algorithm. Specifically, the root of the clique tree, $C(\Omega)$, first collect information iteratively from other nodes. After all the information is collected, it is distributed downwards along the tree to finish the information sharing.

Information collection. For $i = 1, 2, \ldots, 8$, let $\boldsymbol{\phi}_i(A) = \Pr(\boldsymbol{Y}(A) \mid C(A_p) = c_i)$ be the prior predictive distribution of $\boldsymbol{Y}(A)$ evaluated at the observed counts given $C(A_p) = c_i$. When $A \in \mathcal{I} \backslash \{\Omega\}$, it is given by

$$
\begin{aligned}
\boldsymbol{\phi}_i(A) &:= \Pr(\boldsymbol{Y}(A) \mid C(A_p) = c_i) \\
&= \sum_{1 \leqslant i' \leqslant 8} \Pr(\boldsymbol{Y}(A) \mid C(A) = c_{i'}) \Pr(C(A) = c_{i'} \mid C(A_p) = c_i) \\
&= \sum_{1 \leqslant i' \leqslant 8} \boldsymbol{\xi}_{ii'}(A) M_{c_{i'1}}(A) \Pr(\boldsymbol{Y}(A_l) \mid C(A) = c_{i'}) \Pr(\boldsymbol{Y}(A_r) \mid C(A) = c_{i'}) \\
&= \sum_{1 \leqslant i' \leqslant 8} \boldsymbol{\xi}_{ii'}(A) M_{c_{i'1}}(A) \boldsymbol{\phi}_{i'}(A_l) \boldsymbol{\phi}_{i'}(A_r).
\end{aligned}
\tag{3.12}
$$

When A has no children in $\mathcal{I}$, by definition $\boldsymbol{\phi}_i(A) = 1$ for all i. When $A = \Omega$, $C(A)$ does not have a parent in the clique tree. To simplify the notation, we can introduce an imaginary node Ω_p that serves as Ω's parent. Without loss of generality, we let Ω be the left child of Ω_p and let Ω_s with $\boldsymbol{\phi}_i(\Omega_s) = 1$ be Ω's "imaginary" sibling. By setting $\Pr(C(\Omega) = c_{i'} \mid C(\Omega_p) = c_i)$ to the corresponding marginals of $C(\Omega)$ for

$c_i, c_{i'} \in \boldsymbol{c}$, we can keep the formulation in (3.12).

As a byproduct of the inference algorithm, $\phi_1(\Omega)$ gives the marginal likelihood of the data, which can be used to guide the selection of hyper-parameters (Section 3.2.6). Moreover, since the marginal likelihood is available, variable selection can be achieved by putting spike-and-slab priors on the regression coefficients (details are in Appendix A.3).

Information distribution. By Bayes' theorem, the posterior transition probability

$$\begin{aligned}
\tilde{\boldsymbol{\xi}}_{ii'}(A) &:= \Pr(C(A) = c_{i'} \mid C(A_p) = c_i, \boldsymbol{Y}) \\
&= \frac{\Pr(C(A) = c_{i'}, \boldsymbol{Y}(\Omega) \mid C(A_p) = c_i)}{\Pr(\boldsymbol{Y}(\Omega) \mid C(A_p) = c_i)} \\
&= \frac{\Pr(C(A) = c_i' \mid C(A_p) = c_i) \Pr(\boldsymbol{Y}(A) \mid C(A) = c_i')}{\Pr(\boldsymbol{Y}(A) \mid C(A_p) = c_i)} \\
&= \frac{\boldsymbol{\xi}_{ii'}(A) M_{c_{i'1}}(A) \boldsymbol{\phi}_{i'}(A_l) \boldsymbol{\phi}_{i'}(A_r)}{\boldsymbol{\phi}_i(A)}.
\end{aligned} \tag{3.13}$$

Let $\tilde{\boldsymbol{\xi}}(A)$ be the matrix of the posterior transition probabilities of $C(A)$ given $C(A_p)$. Based on $\{\tilde{\boldsymbol{\xi}}(A) : A \in \mathcal{I}\}$, it is easy to compute the PMAP on each node $A \in \mathcal{I}$. Specifically, starting from the root of the tree, each row of $\tilde{\boldsymbol{\xi}}(\Omega)$ represents the posterior marginals of the clique $C(\Omega)$. Thus

$$\Pr(C(\Omega) = c_{i'} \mid \boldsymbol{Y}) = \sum_{1 \leqslant i \leqslant 8} \tilde{\boldsymbol{\xi}}_{ii'}(\Omega)/8 \tag{3.14}$$

for $i' = 1, 2, \ldots, 8$. By marginalization, we can get PMAP(Ω). The PMAPs on other nodes can be computed by induction. For example, given the clique marginals of $C(A_p)$, we can compute the clique marginals of $C(A)$

$$\Pr(C(A) = c_{i'} \mid \boldsymbol{Y}) = \sum_{1 \leqslant i \leqslant 8} \Pr(C(A_p) = c_i \mid \boldsymbol{Y}) \tilde{\boldsymbol{\xi}}_{ii'}(A) \tag{3.15}$$

for $i' = 1, 2, \ldots, 8$, from which the PMAP on A can be obtained by marginalization. We can also compute the **posterior joint alternative probability** (PJAP) that captures the empirical evidence against the global null. Formally,

$$\text{PJAP} = 1 - \Pr(S(A) = 0, A \in \mathcal{T} \mid \boldsymbol{Y}) = 1 - \prod_{A \in \mathcal{I}} \tilde{\boldsymbol{\xi}}_{11}(A). \tag{3.16}$$

Note that the above steps are a variant of the junction tree algorithm (Lauritzen and Spiegelhalter, 1988) that efficiently calculate the marginals in a graphical model. Directly applying the standard junction tree algorithm on $\mathcal{J}$ outputs the PMAPs, however, it does not allow efficient computation of the PJAP. Algorithm 1 summarizes the entire inference recipe for testing the existence of cross-sample differences under BGCR.

3.2.5 Decision making

The PJAP and the PMAPs provide the bases of making decisions about the original and the node-specific hypotheses along the tree. For the original testing problem, we reject the null if PJAP > 0.5, which corresponds to the Bayes optimal decision rule under the simple 0-1 loss. Similarly, we reject $H_0(A)$ and call A a significant node if PMAP$(A) > L$ for some threshold $0 < L < 1$. For example, $L = 0.5$ is recommended by Barbieri et al. (2004). Details on decision making can be found in Appendix A.2. It is worth noting that by reporting the significant nodes and marking them on the phylogenetic tree (or simply mark the PMAPs along the tree without explicitly providing a decision), we have a natural way to characterize and visualize the cross-group differences that offers more insights than merely providing a decision about a test on the original null. For example, it sheds light on the set of microbes most relevant to the cross-group differences.

Algorithm 1 BGCR for comparing microbiome composition

Construct the clique tree $\mathcal{J}$. ▷ Preprocessing

for A in $\mathcal{I}$ **do**
Compute the marginal likelihoods of the local test on A.
Compute the prior transition matrix $\boldsymbol{\xi}(A)$ on the clique of A.
Compute $D(A)$ — the depth of A defined as the number of edges from A to the root of $\mathcal{T}$.
end for

for d in $\max_{A\in\mathcal{I}}\{D(A)\} : 0$ **do** ▷ Recursive information collection
for A with $D(A) = d$ **do**
if $C(A)$ has no children in $\mathcal{J}$ **then**
Let $\boldsymbol{\phi}_i(A_l) = \boldsymbol{\phi}_i(A_r) = 1$, $i = 1, 2, \ldots, 8$.
else
Compute $\boldsymbol{\phi}_i(A) = \sum_{1\leqslant i'\leqslant 8} \boldsymbol{\xi}_{ii'}(A) M_{c_{i'1}}(A) \boldsymbol{\phi}_{i'}(A_l) \boldsymbol{\phi}_{i'}(A_r), \quad i = 1, 2, \ldots, 8.$
end if
end for
end for

for A in $\mathcal{I}$ **do** ▷ Information distribution

$$\text{Compute} \quad \tilde{\boldsymbol{\xi}}_{ii'}(A) = \frac{\boldsymbol{\xi}_{ii'}(A) M_{c_{i'1}}(A) \boldsymbol{\phi}_{i'}(A_l) \boldsymbol{\phi}_{i'}(A_r)}{\boldsymbol{\phi}_i(A)}, \quad i, i' = 1, 2, \ldots, 8.$$

end for

Compute the PJAP. ▷ Global information summary

for d in $0 : \max_{A\in\mathcal{I}}\{D(A)\}$ **do** ▷ Local information summary
for A with $D(A) = d$ **do**
if $A = \Omega$ **then**
Let $\Pr(C(A_p) = c_i \mid \boldsymbol{Y}) = 1/8$, i = 1,2,..., 8.
else
Compute $\Pr(C(A) = c_{i'} \mid \boldsymbol{Y}) = \sum_{1\leqslant i\leqslant 8} \Pr(C(A_p) = c_i \mid \boldsymbol{Y}) \tilde{\boldsymbol{\xi}}_{ii'}(A)$, $i' = 1, 2, \ldots, 8$.
Compute PMAP(A) by marginalization.
end if
end for
end for

Report nodes with PMAP(A) $> L$ for some threshold L. ▷ Decision making
Reject the original null if PJAP > 0.5.

3.2.6 Specifications of hyper-parameters

In BGCR, we need to specify $\alpha(A), \tau(A)$ and $\kappa(A)$ in (3.8). For simplicity, we let α, τ, κ be global parameters that are the same for all $A \in \mathcal{I}$.

Choice of α. The **prior joint alternative probability** (PrJAP) can be written as

$$\text{PrJAP} = 1 - \left(\frac{1}{1+\exp(\alpha)}\right)^{|\mathcal{I}|},$$

which is a monotone function of α. We can then choose $\alpha = \alpha_0$ such that the PrJAP is at a desired level such as 0.5.

Choice of κ. τ and κ together control the "stickiness" of the signals, that is, how likely the "chaining" pattern will occur along the phylogenetic tree. To encapsulate the "explaining away" effect, we set $\kappa = 0$ and assume that the signal at A could be "explained away" by the signal at one of its children. Another choice for κ is $\kappa = \tau$, which imposes additive effect of signals at the sibling nodes on the parent.

Choice of τ. Given $\alpha = \alpha_0, \kappa = 0$, we use an empirical Bayes procedure to choose the value of τ. Given τ, the marginal likelihood is given by $\boldsymbol{\phi}_1(\Omega)$ as defined in (3.12). We maximize $\boldsymbol{\phi}_1(\Omega)$ over τ to get $\hat{\tau}$ as the prior choice for τ. In search of τ, we focus on the interval $[0, \tau_{max})$, where τ_{max} is some predetermined upper bound on τ. One way to choose τ_{max} is based on the *prior marginal alternative probability* (PrMAP) on each node, which is a monotonically increasing function of τ. Therefore, the total number of nodes with PrMAP exceeding the threshold L also increases with τ. We can preselect an upper bound of the total prior expected number of true alternatives and solve for τ_{max} accordingly. In practice, when the number of OTUs is not very large ($\leqslant 150$), we find that setting $\tau_{max} = 6$ usually works well (this corresponds to setting the prior expected sum of PrMAPs equal to 2 with 100 total OTUs). When $\tau = 0$, BGCR incorporates BCR as a special case.

3.3 Numerical examples

3.3.1 Evaluating the performance of BGCR

We first carry out three simulation studies to illustrate the performance of BGCR. For these simulations, several synthetic datasets are obtained based on the July 29, 2016 version of the fecal data of the AG project introduced in Chapter 2.

Although the number of OTUs in the study is large, many cells in the OTU table contain zero or very few counts. For illustration purpose, we focus on the top 50, 75, and 100 OTUs with the largest overall counts across samples. Using different numbers of OTUs allow us to evaluate the inference of increasing dimensionality on the performance of different tests. Note that even with these filtered top OTUs, the data still demonstrate the four challenging features of the OTU counts listed in Section 4.1. We further narrow down the samples to 561 middle-aged (people in their 30s, 40s and 50s) male Caucasian participants from the west census region to reduce the large variation across different simulation rounds (we will use the full dataset in our later data analysis). Reducing the sample size in the simulation is also necessary due to the speed of the competitors. In each simulation, we randomly divide the data into two roughly equal-sized groups to create the data under the null. The data under the alternative is generated under three scenarios:

I. **Cross-group difference exists at a single OTU.** This scenario is also considered in Tang et al. (2018). In each round of the simulation, we randomly select an OTU and increase its count in the second group by a given percentage p. This induces cross-group differences on the parent node of the selected OTU. In this case, the differences are considered local on the tree. We consider $M = 50, 75$, and 100 with $p = 100\%, 150\%$, and 250% (To rule out obvious simulations with either too weak or too strong signals, we place additional constraints in the random selection. Specifically, we only select OTUs with

sample means in the middle 80% range of all the OTU sample means).

II. **Cross-group difference exists at multiple OTUs.** Similar to the first scenario, *eight* OTUs are randomly selected and their counts are increased by a given percentage p of the subjects in the second group to create cross-group differences that are more global on the tree. We consider $M = 50, 75$, and 100 with $p = 50\%, 100\%$, and 150%.

III. **Cross-group difference exists at a chain of nodes in the phylogenetic tree.** Cross-group differences of OTU compositions often cluster into chains as shown in Section 4.4 and Tang et al. (2018). To see how the graphical structure in BGCR helps increase the power of the test in this situation, we consider a case in which cross-group differences are present in a fixed chain of nodes in the phylogenetic tree. In particular, we focus on the top 100 OTUs ($M = 100$) with their phylogenetic tree shown in Figure 3.3. Consider a chain of three nodes in the tree as shown in red in Figure 3.3, we create cross-group differences at these nodes by systematically modifying the counts of their descendant OTUs. Specifically, we increase the counts of OTU 1, 2 and 3 (marked in Figure 3.3) of all the subjects in the second group by $0.33p, 0.67p$ and p percent, respectively. We consider the case with $p = 75\%, 100\%$, and 125%.

In each scenario, we carry out 3000 rounds of simulations. For BGCR, we let $\gamma(A) \sim \mathrm{N}(0, 10)$ and $\log_{10} \nu(A) \sim \mathrm{Unif}(-1, 4)$ independently for $A \in \mathcal{I}$. Hyper parameters are chosen according to Section Figure 3.2.6 with PrJAP = 0.5, $\kappa = 0$, and $\tau_{max} = 6$. We compare BGCR with DTM using the maximum of the single node statistic (DTM-1), using the maximum of the triplet statistic (DTM-3) (Tang et al., 2018) and the DM test (La Rosa et al., 2012a).

The ROC plots under the three simulation scenarios with $M = 100$ are shown in Figure 3.4. Similar results under Scenario II and III with $M = 50$ and $M = 75$ and

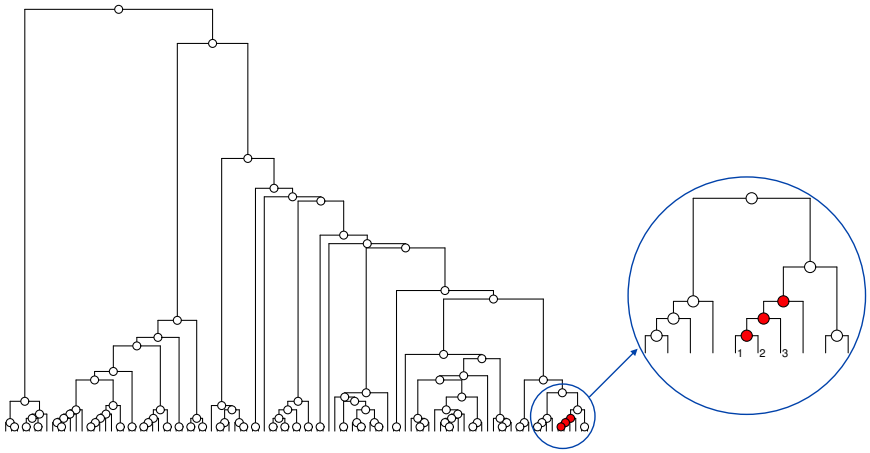

FIGURE 3.3: The phylogenetic tree of the top 100 OTUs with the largest overall counts. Each circle represents an internal node; each tip represents an OTU. The chain of nodes with designed cross-group differences in simulation III is marked red. The zoom-in plot gives a better view of these nodes and their descendant OTUs.

shown in Figure A.1 and Figure A.2 in Appendix A.4. Overall, BGCR outperforms the competitors in all these scenarios. In general, the three methods that incorporate phylogenetic information perform better than the DM test. In Scenario I and II, Figure 3.4 and Figure A.1 show that DTM-1 and DTM-3 work better when M is small. Compared with BGCR, these tests perform worse with the increase of . This is due to the inherent features of the microbiome data. When M is small, the samples typically contain large counts of many OTUs included in the study and are likely to be described by the DM-type models. When M becomes larger, the OTU table gets sparser, making it harder for DTM-1 and DTM-3 to estimate the right dispersion parameters. In contrast, the dispersion parameters are integrated out rather than estimated in BGCR.

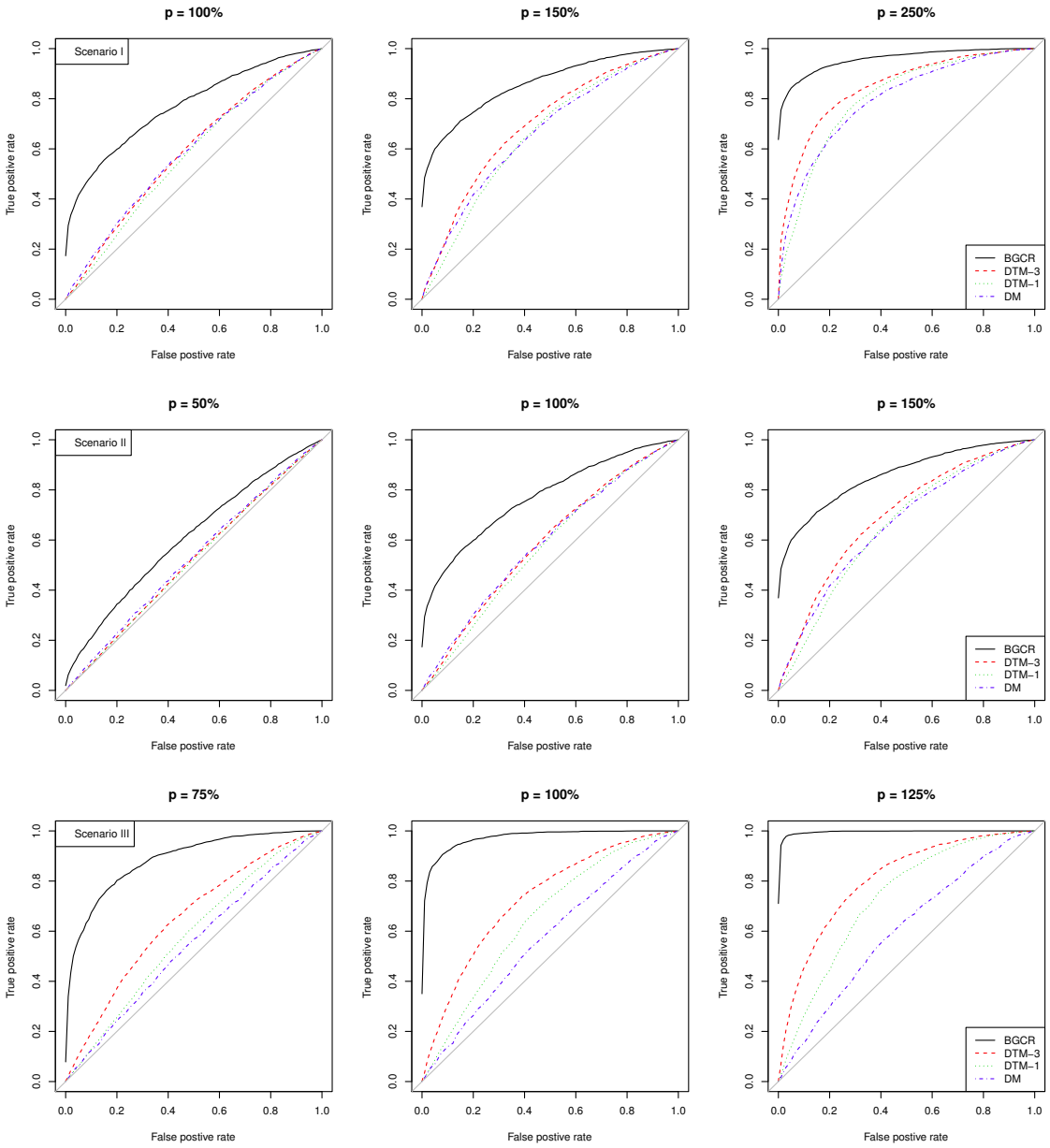

FIGURE 3.4: ROC curves for Scenario I (top row), II (middle row) and III (bottom row) with $M = 100$. The columns are indicated by the percent of counts increased in the second group (p).

To illustrate how borrowing information among neighboring nodes helps increase the power of the node-specific tests, we compare BGCR with BCR in Scenario III. Under the null, the estimated γ by the empirical Bayes procedure in BGCR are zero in all but a few of the simulations (Figure A.3 left). Therefore, BGCR essentially degenerates to BCR and does not increase the chance of false positives (Figure A.3 middle and right). Under the alternative, by introducing positive dependencies among the local tests, BGCR obtains larger PJAPs and thus provides a better chance to correctly reject the null. To see how this would influence the decision procedure, we show the proportion of tests that reject the null under different decision thresholds L in Figure A.4 in Appendix A.4. When $0.5 < L < 0.9$, the gain by introducing the dependencies among tests is remarkable. Figure A.5 in Appendix A.4 shows the estimated γ's under the alternatives. When the cross-group difference increases, the chaining pattern gets stronger, causing the estimated γ to increase. Figure A.6 in Appendix A.4 shows the corresponding PJAPs of the two models under comparison.

3.3.2 Visualizing the cross-group differences with BGCR

In many applications, characterizing and visualizing the cross-group difference is often of more interests than merely testing its existence. As mentioned in Section 3.2.5, this can be achieved by marking the PMAPs along the phylogenetic tree. As an example, consider a specific round of simulation in Scenario III. In this simulation, $\text{PJAP}_{\text{BGCR}} = 0.910$, $\text{PJAP}_{\text{BCR}} = 0.651$, $\hat{\gamma} = 5.521$. Figure 3.5 plots the PMAPs obtained by the two models along the phylogenetic tree (only shows the relevant part of the tree). It is clear that BGCR benefits from the graphical structure and reveals the designed chaining pattern of the cross-group differences (see Figure 3.3).

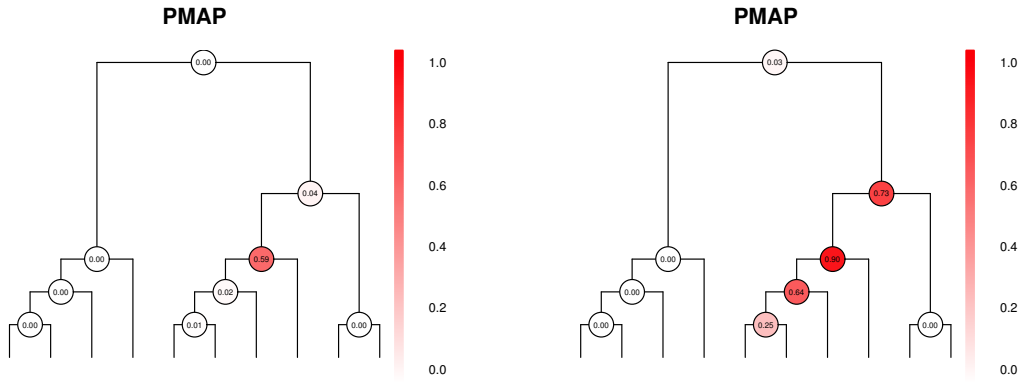

FIGURE 3.5: PMAPs from BCR and BGCR under a specific simulation case in Scenario III. Only nodes in the zoom-in subtree in Figure 3.3 are shown since the PMAPs at other nodes are essentially zero. Left: PMAPs obtained by BCR; Right: PMAPs obtained by BGCR. The number in each node denotes its PMAP. We also color the nodes by their PMAPs to give a better visualization.

3.3.3 Adjusting for covariates

In practice, OTU compositions can depend on various factors. When the data are gathered from observational studies, unadjusted covariates can lead to false positives. Even when the data do arise from randomized experiments, incorporating relevant covariates can improve the statistical power for identifying cross-group differences. We illustrate with another simulation scenario the necessity of incorporating confounders in the testing procedure:

IV. **Cross-group difference exists at a single OTU with an unbalanced confounder.** We start from the same American Gut dataset as in the previous simulation scenarios. Now we randomly select 200 *male* and 200 *female* Caucasians, all middle aged and from the west region (the reason we take a random sample instead of using the full dataset is the large computational burden brought by the competitors). To purposefully make gender an unbalanced confounder, we select an OTU (OTU '4352657', denoted by ω_c) and increase

the counts of this OTU for all male participants by 175%. We then randomly select 160 males and 40 females into the first group and put the rest into the second group to get the data under the null. To simulate data under the alternative, we select another OTU (OTU '4481131', denoted by ω_s) and increase the count of this OTU in the second group by 175%.

We carry out 750 simulations with $M = 50$ and compare BGCR with DTM-3 and DM (we do not simulate 3000 rounds or consider $M = 75$ or 100 since some competitors are too slow). Figure 3.6 shows the histograms of the statistics these methods used for decision making under the null—PJAP for BGCR, p-values (or equivalently, $1 - p$-values) for DTM-3 and DM. From Figure 3.6 (ii)-(iv), it is clear that all three methods fail in controlling false positives without adjusting for the "gender" of the participants. After incorporating "gender" in each node-specific regression, BGCR is able to keep the PJAPs under the null at a reasonable level (Figure 3.6 (i)).

Adjusting for confounders is also important in the characterization of the cross-group differences. For example, consider a specific round of simulation in Scenario IV. Figure 3.7 marks the PMAPs reported by BGCR on the phylogenetic tree under the null and the alternative, with or without "gender" included. Under the null, ignoring "gender" leads to a false discovery at the parent of ω_c and also a false rejection of the hypothesis that no cross-group difference exists at threshold 0.95. Under the alternative, the PJAPs reported by BGCR are closed to 1 with or without adjusting for "gender", suggesting that incorporating "gender" in the model is likely to keep the decision about the original hypothesis unchanged. However, failing to adjust for "gender" contaminates the pattern of the cross-group differences (Figure 3.7 (iii)-(iv)).

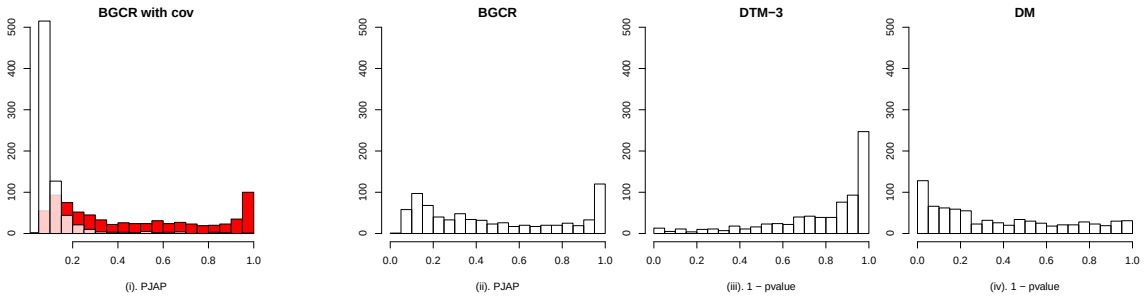

FIGURE 3.6: Histograms of the statistics used for decision making under the null. Plot (i) shows the histogram of PJAPs under the null (white) and under the alternative (red) with "gender" adjusted. The other three plots present the histograms of the decision-making statistics of the three models under comparison under the null without adjusting for "gender".

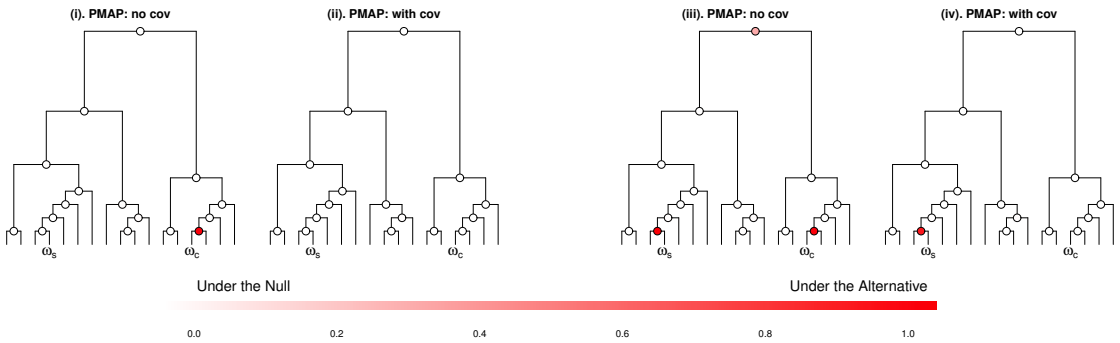

FIGURE 3.7: PMAPs of a specific simulation in Scenario IV. (i) and (ii): PMAPs under the null; (iii) and (iv): PMAPs under the alternative (only a subtree is shown since PMAPs on other nodes are essentially zero). The nodes are colored by their PMAPs. Under the alternative, true cross-group differences are only present at the parent of ω_s. A false discovery is made at the parent of ω_c when "gender" is not adjusted for.

3.4 Application to the American Gut project data

3.4.1 Cross-group comparison of OTU compositions

In this section, we apply BGCR to the same dataset from the AG project as in Section 4.3. The AG project collects many binary or ordered categorical covariates of the participants that can be used to divide them into groups. In particular, we focus on comparing the OTU compositions of groups of participants categorized by their dietary habits, which can be characterized by their frequencies of consuming different kinds of food. We consider eight types of food: alcohol, diary, fruit, meat/egg, seafood, sweets, vegetable and whole grain. For each kind of food, the dataset records each participant's frequency of consumption per week as a categorical variable with five categories: Never, Occasionally (1-2 times per week), Rarely (less than once per week), Regularly (3-5 times per week) and Daily. We dichotomize the categories based on whether the food is consumed less than 3 times per week and use this binary variable to divide the participants into two groups.

In our analysis, we only use the top 100 OTUs with the largest overall counts across samples. This choice is made mainly for quality control. For OTUs containing nil or only very few counts over all samples, their counts are largely influenced by sequencing errors and are highly unreliable. Using these OTUs is likely to contaminate the analysis by bringing in pure noises that are hard to quantify. We note that prescreening the OTUs to drop those who do not have substantial counts in any samples is a common recommended practice in the literature (see for example Wu et al. (2011); Tang et al. (2018)).

Besides diet, there are other possible factors that may influence the OTU composition. Including these factors in the model could help reduce the chance of false discoveries and improve the power of the inference. In our analysis, we consider five such covariates: the participant's biological sex, has the participant been diagnosed

with diabetes, does the participant have IBD (inflammatory bowel disease), has the participant used antibiotic in the past year, and does the participant consume probiotic less than 3 times per week. All five covariates are binary. In addition to adjusting for these non-dietary covariates when investigating a certain kind of food, dietary habits of other foods are also natural candidates for confounders to be included in the analysis, especially when they are unbalanced across the two groups. For example, Figure 3.8 shows the unbalancedness of vegetable and grain consumption of participants across the two groups defined by fruit consumption. Since the covariate information in the study is collected by asking the participants to take a non-compulsory questionnaire, some covariates are missing for certain participants. For simplicity, we ignore all participants with missing values in any variable mentioned above and perform a complete-case analysis. Moreover, we restrict our samples to participants with age in range $20 \sim 69$ and BMI $18.5 \sim 30$ to disregard potential outliers. The resulting sample size in each comparison is shown in Table 3.2.

For each type of food, we use BGCR to compare the group-level OTU composition across the two resulting groups. In each comparison, we consider BGCR with different sets of covariates adjusted—first no covariate, then the five non-dietary covariates, and finally the non-dietary covariates as well as dietary habits on other foods. The $(1 - \text{PJAP})$'s reported by BGCR under each comparison are shown in Table 3.2.

When all the covariates are adjusted, the test rejects the global null at threshold $L = 0.5$ when the grouping food is fruit, seafood, vegetable, or whole grain. This is consistent with the findings that dietary habits are closely related to gut microbiome compositions in general (Graf et al., 2015; Singh et al., 2017; Rothschild et al., 2018; Tang et al., 2018).

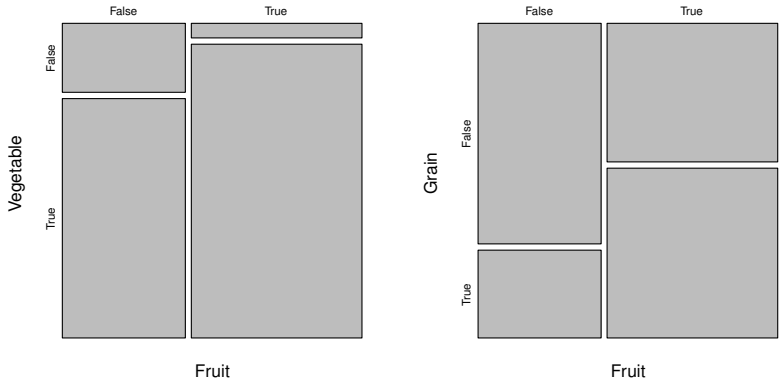

FIGURE 3.8: Unbalancedness of vegetable and grain consumption in the two groups defined by fruit consumption. The area of each block represents the proportion of the counts in the corresponding cell of the 2×2 contingency table.

Table 3.2: (1−PJAP)'s for testing OTU compositions of different dietary habits. In the last column, red cells indicate rejections of the global null under threshold $L = 0.5$. The estimated $\hat{\tau}$ when all the covariates are adjusted is shown in the parentheses in the last column.

Diet	Group size		1−PJAP		
	False	True	No cov	Non-diet cov	All cov ($\hat{\tau}$)
Alcohol	2099	995	0.134	0.345	0.555 (0)
Diary	1733	1361	1.27×10^{-2}	6.97×10^{-2}	0.517 (0)
Fruit	1296	1798	6.86×10^{-7}	1.29×10^{-8}	0.259 (3.93)
Meat/Egg	798	2296	0.867	0.857	0.757 (0)
Seafood	2637	457	1.99×10^{-2}	9.11×10^{-3}	0.166 (0)
Sweets	2005	1089	0.200	0.347	0.912 (0)
Vegetable	375	2719	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$ (5.13)
Whole grain	1735	1359	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	3.85×10^{-7} (5.63)

Our analysis also shows that adjusting for relevant covariates, which is missing in previous analyses, can significantly change the testing results and can reverse the testing decision on the null hypothesis. For example, when no covariate or only non-dietary covariates are adjusted, the null of no cross-group difference is likely to be rejected for diary and sweets, which leads to the same conclusion as in Tang et al. (2018). However, once other dietary covariates are included, these two diets tend to have no significant association with the microbiome composition. Typically, the posterior probability of the null (1 - PJAP) increases when more covariates are incorporated in the model. This reflects the complex nature of the microbial community that it is associated with various factors (Rothschild et al., 2018).

Similar phenomena can be seen on the more refined PMAP scale. For the four comparisons that reject the null, we visualize the cross-group differences by marking the PMAPs on the phylogenetic tree (Figure 3.9). Similar plots of PMAPs under BGCR with no covariate or only non-dietary covariates can be found in Appendix A.4 (Figure A.7 and Figure A.8). These plots show that adjusting for relevant covariates greatly changes the characterization of the cross-group differences. Many significant nodes reported by BGCR with some covariates unadjusted turn out to be false positives when all the covariates are included.

3.4.2 BGCR vs. BCR

In this section, we evaluate the performance of BGCR compared with BCR in our data application. Figure A.9 in Appendix A.4 shows the PMAPs under BCR for the four comparisons that BGCR rejects the null. In BGCR, the global parameter τ controls the level of dependencies among the node-specific tests. We note in Section 3.2.6 that when $\tau = 0$, BGCR degenerates to BCR. This nested feature together with our empirical Bayes strategy to estimate τ allow BGCR to decide whether the dependencies are necessary based on the data—when the estimated $\hat{\tau} = 0$, BGCR

simply performs independent tests without introducing the graphical structure.

Table 3.2 shows $\hat{\tau}$ in each comparison when BGCR with all covariates is used. In all four comparisons that the null is accepted at threshold $L = 0.5$, $\hat{\tau} = 0$. In these cases, no node is significant and the "signals" are trivially independent. For fruit, vegetable and whole grain, $\hat{\tau} > 0$, suggesting that cross-group differences might cluster into chains. To see whether it is necessary to introduce the graphical structure among the node-specific tests, one can test $H_0^g : \tau = 0$ vs $H_1^g : \tau \neq 0$. For example, when the prior $\tau \sim \text{Uniform}(0, 6)$ is used, the Bayes factor comparing H_1^g and H_0^g for fruit, vegetable, seafood and whole grain are 1.022, 0.830, 138.699 and 6.922, respectively. For vegetable and whole grain, the cross-group differences cluster into chains (Figure 3.9) and thus it is necessary to introduce the graphical structure. We can also look at the posterior on τ in the four comparisons that reject the null as shown in Figure 3.10. For vegetable and whole grain, the signals demonstrate a strong chaining pattern and the posteriors on τ concentrate to some positive values. For fruit and seafood, the cross-group differences are local to certain nodes. Therefore, not much information about the dependencies of the node-specific tests is provided by the data and the posteriors on τ are very similar to the prior.

3.5 Concluding remarks

We have introduced a Bayesian framework for testing the existence of cross-group differences in the OTU composition. The original testing problem is transformed into a series of dependent node-specific tests linked together by the phylogenetic tree. A key feature of the Bayesian formulation is that the dispersion parameters are integrated out, providing more robust inference than the methods that use point estimates of the within-group variation. We use numerical integration to evaluate the Bayes factors, circumventing Monte Carlo simulations to give fast inference. By introducing dependencies among the local tests, information sharing is allowed

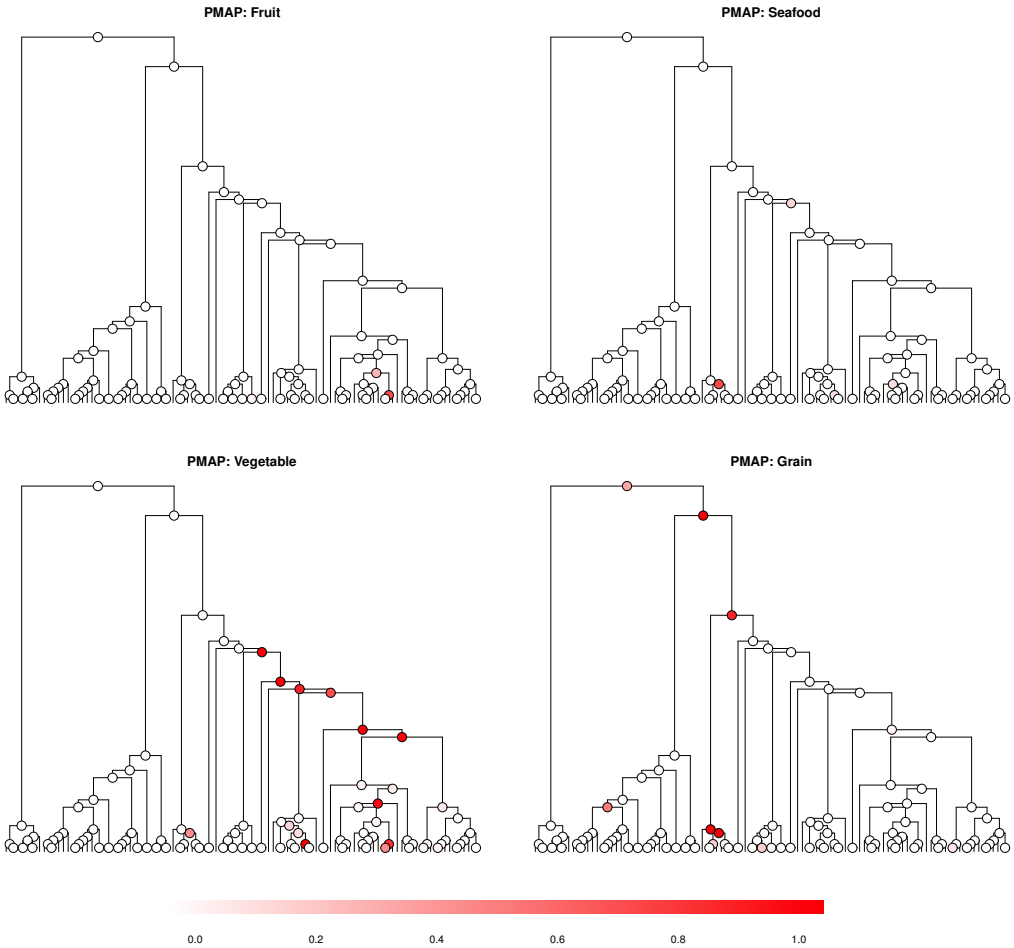

FIGURE 3.9: PMAPs for the four comparisons that reject the global null. The nodes are colored by PMAPs reported by BGCR with both non-dietary covariates and dietary covariates adjusted.

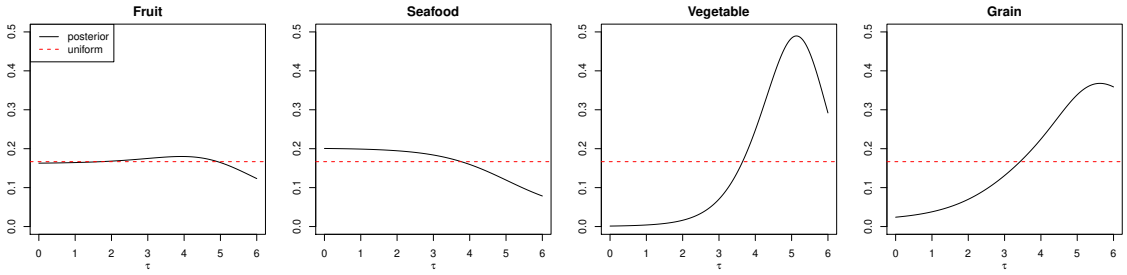

FIGURE 3.10: Posteriors on τ with the uniform prior. Shown for the four comparisons that reject the null at threshold $L = 0.5$.

among neighboring nodes. This further improves the power of the test when the cross-group differences cluster into chains along the tree, which is often observed in practice. We derive an exact message passing algorithm that is scalable to the size of the tree to carry out inference under this information borrowing.

Because microbiome data are often collected by observational studies, it is important to adjust for possible confounders in testing. This is crucial in reducing the chance of false discoveries and the residual overdispersion. Our model achieves this goal by incorporating the covariates via a regression model, which adds little computational burdens to the Bayesian testing framework. A relevant problem is to select covariates in the testing scenario. Generally, this must be done with caution since incorporating standard variable selection procedures can substantially affect or even invalidate the meaning of the testing result on the two-group difference due to possible confounding. Readers may refer to Appendix A.3 for more details on this along with some numerical examples.

Besides testing, our model gives a full probabilistic characterization of the cross-group differences that can be naturally visualized. Note that although we focus on two group comparisons in this chapter, the proposed method can be generalized to multiple group settings with little extra effort.

4

Dirichlet-tree Multinomial Mixtures

4.1 Introduction

The microbiome is the collective genomes of all microbes that inhabit a specific environment, be it an animal, a plant, or a tissue. In particular, the human microbiome is considered as a counterpart or an extension to the human genome (Grice and Segre, 2012), containing 100-fold more unique genes than the latter (Qin et al., 2010). The human microbiome has significant influences on various aspects of our physiology (Turnbaugh et al., 2009; Qin et al., 2012; Karlsson et al., 2013) and is suggested as a way towards precision medicine (Kuntz and Gilbert, 2017). The development of next-generation sequencing strategies enables us to profile the microbiome fast and economically, through either amplicon sequencing on target genes (usually the 16S ribosomal RNA gene) or shotgun sequencing on the entire microbial genome. In this work, we focus on datasets obtained from amplicon sequencing studies. Traditionally, the sequencing reads are sent to preprocessing pipelines such as QIIME (Caporaso et al., 2010) to construct clusters named operational taxonomic units (OTUs) based on certain predefined similarity threshold (typically 97%). In contrast, more recently

developed pipelines such as DADA2 (Callahan et al., 2016) directly resolve amplicon sequence variants (ASVs), which is shown to outperform OTUs in terms of accuracy and interpretability (Callahan et al., 2017). OTUs and ASVs serve as the unit for the downstream analyses and provide the same interface: each sample is a vector of counts on a list of units (OTUs or ASVs), representing the composition of the underlying community. The methodology developed in this work applies to both OTUs and ASVs, we thus use the customary "OTU" to refer to the unit.

Typical tasks of microbiome research include characterizing the microbiome composition of an individual and study its relationship with various covariates, such as the host's dietary patterns or disease status. Due to the considerable variability of human microbiome compositions, it is hard to complete these tasks directly at the individual level (Ding and Schloss, 2014). A popular way to overcome this difficulty is to first group individual samples into clusters and perform downstream analyses at the cluster level. For example, in the context of the human gut microbiome, these clusters are referred to as "enterotypes" (Arumugam et al., 2011; Costea et al., 2018), which are shown to be associated with long-term dietary habits, obesity, and Crohn's disease (Wu et al., 2011; Holmes et al., 2012; Quince et al., 2013). Nevertheless, clustering microbiome samples is challenging. Off-the-shelf machine learning algorithms such as k-means, Partitioning Around Medoids (PAM) and hierarchical clustering that are distance-based are not satisfactory when applied to microbiome data. To get a partition of the samples into clusters, these methods entail a predetermined number of clusters or cut-offs, which is typically hard to specify in microbiome settings. In addition, popular distance metrics for microbiome compositional data such as the Bray-Curtis dissimilarity and the Unifrac distances (Lozupone and Knight, 2005) can induce improper weights to different OTUs while clustering. Koren et al. (2013) show that different methodologies to select the number of clusters can yield inconsistent clustering results and that these algorithms are sensitive to the distance

metrics chosen.

Alternatively, one can achieve clustering through building a probabilistic model for the microbiome count data with a latent grouping structure, such as the mixture model. The key to this thread is a good generative model for the OTU counts: firstly, the model needs to be flexible enough to capture the idiosyncrasy of each cluster and allow the complex variations of samples therein; secondly, it should borrow information across all samples to accurately infer the shared portion of all clusters; moreover, since biological information of the OTUs is available, it is desirable for the model to incorporate such information. By far, the most popular model-based clustering method designed for microbiome data is Dirichlet multinomial mixtures (DMM) (Holmes et al., 2012), which adopts a multinomial sampling scheme and generates the sample-specific multinomial parameters from a finite mixture of Dirichlet distributions. Unfortunately, DMM does not satisfy the above desiderata.

We illustrate some limitations of DMM with a simple example with six OTUs. Suppose that the microbiome population has three equal sized clusters. We first assume DMM is the true data generating process and denote the three Dirichlet components as $\text{Dir}(\boldsymbol{\alpha}_k)$, $k = 1, 2, 3$. Let $\boldsymbol{\alpha}_1 = (10, 2, 4, 4, 2, 2) \cdot \alpha_0$, $\boldsymbol{\alpha}_2 = (6, 6, 4, 4, 2, 2) \cdot \alpha_0$, $\boldsymbol{\alpha}_3 = (2, 10, 4, 4, 2, 2) \cdot \alpha_0$, respectively. 4.1 (a) and (c) show the 2.5% and 97.5% percentiles of the marginals of $\text{Dir}(\boldsymbol{\alpha}_k)$ for $\alpha_0 = 10$ and $\alpha_0 = 0.25$, $k = 1, 2, 3$. Note that the single parameter α_0 governs the dispersion level of all the categories simultaneously. Therefore, although the Dirichlet distribution is flexible in capturing each cluster centroid, it is restrictive in modeling the systematic variations among samples in the same cluster around the centroid. Consequently, when DMM is applied to data with true marginals on the multinomial parameters as shown in 4.1 (b), where different categories have different levels of dispersions, α_0 has to be estimated small to accommodate to the large dispersion of the last 4 categories, essentially washing away the strong signals contained in the first 2 categories and pooling all the samples

to the same cluster. Unfortunately, this situation is not uncommon for microbiome data. In this example, the clusters are only different in the compositions of OTU 1 and 2. However, DMM ignores this common feature across clusters and overfits the data by introducing cluster-specific parameters for the compositions of the last 4 OTUs. This problem is amplified in real microbiome applications, where the number of OTUs is typically large and the overfitting can be severe when the clustering is relevant to only a portion of OTUs.

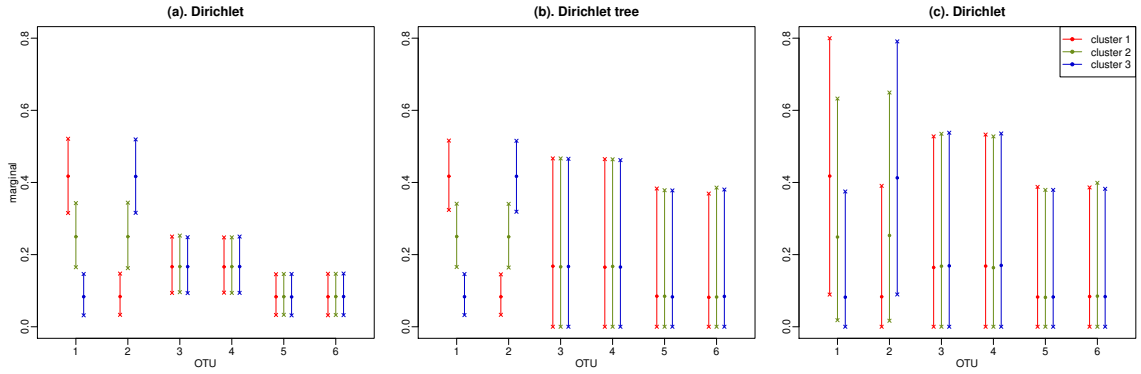

FIGURE 4.1: 2.5% and 97.5% percentiles of the marginals for each category (plots are obtained based on simulation). (a): the truth is the Dirichlet distribution, with small dispersion; (b): the truth is the Dirichlet-tree distribution (introduced in 4.2); (c): the truth is the Dirichlet distribution, with large dispersion.

In this work, we propose a method, called Dirichlet-tree multinomial mixtures (DTMM), that performs model-based clustering of the microbial compositional data. Similar to DMM, our method applies the mixture model to achieve clustering under the Bayesian inference framework. Unlike DMM, we consider a more flexible mixing kernel than the Dirichlet distribution that (i) provides a more flexible covariance structure among OTU counts within a cluster and (ii) offers a solution to identify and efficiently infer the common parts across clusters. By directly bringing in a phylogenetic tree of the OTUs that summarizes their evolutionary relationship, DTMM adopts the Dirichlet-tree distribution (DT) (Dennis, 1991) as the mixture

component that sequentially generates the multinomial parameters through a series a beta distributions at the internal nodes of the phylogenetic tree. On one hand, DT introduces a set of dispersion parameters through these beta distributions that allow different variation levels among the OTUs, offering a more flexible covariance structure than the Dirichlet distribution. 4.1 (b) gives an example of the marginals generated from three DT under a specific tree. On the other hand, through allowing some of these beta distributions to be shared across clusters, cluster-specific parameters are only introduced when necessary and the common portion of the clusters can be inferred accurately.

Using phylogenetic information to guide inference is a popular direction in recent microbiome research. For example, Wang and Zhao (2017), Tang et al. (2018) and Mao et al. (2020) use a similar DT-multinomial framework in various model-based microbiome analyses. Silverman et al. (2017) propose a phylogenetic-based log-ratio transform on the compositional data to a space where classical statistical tools can be applied. Other than the Dirichlet-based models, log-ratio based models such as the logistic-normal model (Atchison and Shen, 1980) are also frequently used in microbiome studies (see for example Xia et al. (2013) and Li et al. (2018)). Although offering a more flexible covariance structure than the Dirichlet distribution, these models incur challenges in inferring the high-dimensional covariance structure. As we shall see, the DT model strikes a balance between the simple but restrictive Dirichlet model and the more flexible log-ratio based models: in addition to the aforementioned modeling benefits, the tree-based decomposition allows a divide-and-conquer strategy to perform inference, in a way similar to the posterior inference of Pólya tree type models in the Bayesian nonparametric literature (Lavine et al., 1992, 1994). Instead of having to resort to optimization techniques to do inference as in DMM, fully Bayesian inference becomes practical.

The rest of the paper is organized as follows. Section 4.2 introduces a phyloge-

netic tree-based decomposition of the multinomial counts and proposes DTMM for clustering OTU counts based on this decomposition. Section 4.3 performs a series of representative numerical experiments to evaluate the performance of DTMM and validate the resulting clusters. Section 4.4 applies DTMM to a specific dataset from the American Gut project to find enterotypes of samples that are diagnosed with IBD or diabetes. Section 4.5 concludes.

4.2 Method

4.2.1 DM and DMM

Consider a microbiome dataset with OTU counts of n samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$. Each sample is a vector of counts of the M OTUs in the study. We denote the OTUs by $\Omega = \{\text{OTU}_2, \text{OTU}_2, \ldots, \text{OTU}_M\} = \{\omega_1, \omega_2, \ldots, \omega_M\}$. Let the i-th sample and the total counts in that sample be $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iM})$ and $N_i = \sum_{j=1}^{M} y_{ij}$, where y_{ij} is the count of OTU j. The samples can be stacked into an OTU table denoted by $\boldsymbol{Y}$, as shown in Table 4.1. In this chapter, we shall treat the total counts N_i's as given since they are artificial quantities that depend on the sequencing depth rather than reflecting the true abundance of the OTUs in the sampled environment. This leads to the generative Dirichlet-multinomial model (DM) that takes the multinomial sampling scheme coupled with a Dirichlet prior on the multinomial parameters (Knights et al., 2011; La Rosa et al., 2012b):

$$\begin{aligned} \boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i &\overset{\text{ind}}{\sim} \text{Multi}(N_i, \boldsymbol{p}_i) \\ \boldsymbol{p}_i \mid \boldsymbol{\alpha} &\overset{\text{iid}}{\sim} \text{Dir}(\boldsymbol{\alpha}), \end{aligned} \tag{4.1}$$

where $\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iM})$, p_{ij} is the probability that a count in sample i belongs to OTU j, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ with $\alpha_j > 0$ for $j = 1, \ldots, M$.

Viewing each sample as randomly drawn from an underlying "community" characterized by its multinomial parameter (Holmes et al., 2012), DM models all the

"communities" as realizations of a single "metacommunity" governed by $\boldsymbol{\alpha}$. Holmes et al. (2012) extend DM to Dirichlet multinomial mixtures (DMM) by replacing the single Dirichlet prior in DM by a finite mixture of K Dirichlets to allow clustering of the samples into different "metacommunities":

$$\begin{aligned} \boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i &\overset{\text{ind}}{\sim} \text{Multi}(N_i, \boldsymbol{p}_i) \\ \boldsymbol{p}_i \mid \boldsymbol{\pi}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K &\overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \text{Dir}(\boldsymbol{\alpha}_k) \\ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim \text{Dir}(\boldsymbol{b}_0), \end{aligned} \tag{4.2}$$

where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kM})$, $\boldsymbol{\pi}$ the weights of the "metacommunities" with $\sum_{k=1}^{K} \pi_k = 1, \pi_k \geqslant 0$ for $k = 1, \ldots, K$. In DMM, each sample is viewed as a draw from a unique "community" that is itself drawn from one of the K "metacommunities".

As a clustering method, DMM has several limitations. Most importantly, it could not adequately model the within cluster variation of the microbial composition. This can be seen by writing the cluster-specific Dirichlet parameter $\boldsymbol{\alpha}_k$ as $\boldsymbol{\alpha}_k = \alpha_{k0} \cdot \bar{\boldsymbol{\alpha}}_k$, where $\bar{\boldsymbol{\alpha}}_k$ lying in the $(M-1)$-dimensional simplex represents the prior mean of the multinomial probabilities in cluster k, $\alpha_{k0} = \sum_{j=1}^{M} \alpha_j$ determines the within cluster variation of all these probabilities around $\bar{\boldsymbol{\alpha}}_k$ simultaneously. Secondly, the multinomial parameters in DM are modeled independently up to the sum to one constraint (Mosimann, 1962), which is not suitable in the microbiome context since the OTUs are functionally and evolutionarily related. Moreover, although DMM is specified within the Bayesian framework, the posterior inference is performed by optimization through an EM algorithm with Laplace approximations of the marginal likelihoods. When the number of OTUs is moderate, which is typical in microbiome studies, these techniques are numerically unstable and cannot provide reliable uncertainty quantifications.

Table 4.1: An $n \times M$ OTU table.

Sample	ω_1	ω_2	$\cdots$	ω_M	Sum
1	y_{11}	y_{12}	$\cdots$	y_{1M}	N_1
2	y_{21}	y_{22}	$\cdots$	y_{2M}	N_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nM}	N_n

4.2.2 Dirichlet-tree multinomial mixtures

A key feature of microbiome count data is that the categories (OTUs) are functionally and evolutionarily related. Typically, this relationship can be summarized into a rooted phylogenetic tree. In this phylogenetic tree, each internal node can be viewed as a "taxa" that represents the most recent common ancestor of its descendant OTUs. Let $\mathcal{T} = \mathcal{T}(\mathcal{I}, \mathcal{U}; \mathcal{E})$ be a rooted full binary phylogenetic tree over the M OTUs in the study, where $\mathcal{I}$, $\mathcal{U}$ and $\mathcal{E}$ denote the set of internal nodes, leaves and edges of $\mathcal{T}$, respectively. We denote each node $A \in \mathcal{I} \cup \mathcal{U}$ by the set of its descendant OTUs. In particular, $A = \Omega$ denotes the root of $\mathcal{T}$; $A = \{\omega_j\}$ represents the leaf that contains OTU j for $j = 1, \ldots, M$. With our notation, $\mathcal{U} = \{\{\omega\} : \omega \in \Omega\}$. For $A \in \mathcal{I}$, let A_l and A_r be the left and right children of A, respectively. For $A \in \mathcal{I} \cup \mathcal{U} \backslash \{\Omega\}$, let A_p be its parent and A_s be its sibling (i.e., the node in $\mathcal{T}$ that has the same parent as A).

As is Chapter 3, given $\mathcal{T}$, the multinomial likelihood of $\boldsymbol{y}_i$ factorizes into a series of binomial likelihoods at the internal nodes of $\mathcal{T}$:

$$\mathcal{L}_M(\boldsymbol{y}_i \mid \boldsymbol{p}_i) \propto \prod_{\{A : A \in \mathcal{I}\}} \mathcal{L}_B(y_i(A_l) \mid y_i(A), \theta_i(A)), \tag{4.3}$$

where

$$y_i(A) = \sum_{\{j:\omega_j \in A\}} y_{ij}, \quad \theta_i(A) = \frac{\sum_{\{j:\omega_j \in A_l\}} p_{ij}}{\sum_{\{j:\omega_j \in A\}} p_{ij}}, \tag{4.4}$$
$$y_i(A_l) \mid y_i(A), \theta_i(A) \overset{\text{ind}}{\sim} \text{Binom}(y_i(A), \theta_i(A)).$$

Note that for $j = 1, \ldots, M$, there is a unique path $\mathscr{P}^j = A_0^j = \Omega \to A_1^j \to \cdots \to A_{l_j}^j \to \omega_j$ in $\mathcal{T}$ connecting the root with ω_j such that

$$p_{ij} = \prod_{l=0}^{l_j} \theta_i(A_l^j). \tag{4.5}$$

By (4.4) and (4.5), $\boldsymbol{p}_i$ and $\boldsymbol{\theta}_i = \{\theta_i(A) : A \in \mathcal{I}\}$ give two equivalent parameterizations of the distribution of $\boldsymbol{y}_i$. Figure 4.2 gives an example of how a specific multinomial parameter $\boldsymbol{p}_{\text{ex}} = (\frac{5}{12}, \frac{5}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12})$ on 6 OTUs can be generated sequentially along a given tree.

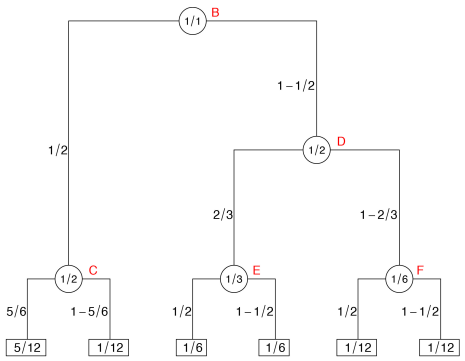

FIGURE 4.2: A tree based generation of $\boldsymbol{p}_{\text{ex}}$.

The factorization in (4.3) provides an orthogonal decomposition of the empirical evidence about $\boldsymbol{p}_i$ into pieces of "local" evidence about $\theta_i(A)$ at $A \in \mathcal{I}$, which suggests a "divide-and-conquer" strategy of doing inference on $\boldsymbol{p}_i$ through learning the "local" probability assignment parameters $\boldsymbol{\theta}_i$. To this end, we take the Bayesian strategy

and put independent beta priors on the binomial parameters:

$$\theta_i(A) \mid \theta(A), \tau(A) \overset{\text{ind}}{\sim} \text{Beta}(\theta(A)\tau(A), (1-\theta(A))\tau(A)), \tag{4.6}$$

where $\theta(A) \in (0,1)$ is the prior mean of $\theta_i(A)$, $\tau(A) > 0$ is a dispersion parameter that controls the variability of $\theta_i(A)$ around its mean. Note that the independent priors on $\boldsymbol{\theta}_i$ together with the relation in (4.5) induce a joint prior on $\boldsymbol{p}_i$, which falls into the family of Dirichlet-tree distributions (DT) (Dennis, 1991). Let $\boldsymbol{\theta} = \{\theta(A) : A \in \mathcal{I}\}$ and $\boldsymbol{\tau} = \{\tau(A) : A \in \mathcal{I}\}$, we shall denote the Dirichlet-tree prior on $\boldsymbol{p}_i$ or $\boldsymbol{\theta}_i$ as

$$\boldsymbol{p}_i \sim \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\tau}) \quad \text{or} \quad \boldsymbol{\theta}_i \sim \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\tau}). \tag{4.7}$$

When $\tau(A) = \tau(A_l) + \tau(A_r)$ for every $A \in \mathcal{I}$ that has non-leaf children, DT degenerates to the Dirichlet distribution. By removing this constraint, DT offers a more flexible way to model the variability of $\boldsymbol{p}_i$ around its cluster centroid.

DT also induces a more flexible covariance structure than the Dirichlet distribution. This can be seen by considering the covariance between any two categories j_1 and j_2 ($j_1 \neq j_2$). Suppose that the first $(L+1)$ nodes in $\mathscr{P}^{j_1}$ and $\mathscr{P}^{j_2}$ are shared and let the shared path be $\Omega = A_0 \to A_1 \to \cdots \to A_L$. It can be shown that (Dennis, 1991)

$$\text{Cov}(p_{ij_1}, p_{ij_2}) = \left[\frac{\tau(A_L)}{\tau(A_L)+1} \prod_{1 \leqslant t \leqslant L} \frac{[a(A_t)+1]\tau(A_{t-1})}{a(A_t)[\tau(A_{t-1})+1]} - 1 \right] \mathbb{E}(p_{ij_1})\mathbb{E}(p_{ij_2}),$$

where $a(A_t) = \theta(A_{t-1})\tau(A_{t-1})$ if A_t is the left child of A_{t-1} and $a(A_t) = (1-\theta(A_{t-1}))\tau(A_{t-1})$ otherwise. In DT, the covariance between categories depends not only on their means and the sum of the pseudo count as in the Dirichlet distribution, but also on the tree structure. This offers a more flexible covariance structure among OTU counts governed by the phylogenetic information. For example, since $a(A_t) < \tau(A_{t-1})$, $[a(A_t)+1]\tau(A_{t-1})/a(A_t)[\tau(A_{t-1})+1] > 1$. Thus p_{ij_1} and p_{ij_2} can

be positively correlated if the share a series of common ancestors in the phylogenetic tree. On the other hand, if p_{ij_1} and p_{ij_2} are "far away" in the phylogenetic tree such that their only common ancestor is Ω, $\mathrm{Cov}(p_{ij_1}, p_{ij_2}) = -\mathbb{E}(p_{ij_1})\mathbb{E}(p_{ij_2})/(\tau(\Omega)+1)$ as in the Dirichlet distribution. When the phylogenetic tree gives decent summaries of the functional relationship among OTUs, this introduces suitable covariance structure among the OTU counts and can improve the inference substantially.

DT has been used for microbiome modelings in various context for different purposes. For example, Wang and Zhao (2017) apply the DT multinomial model to study the association between OTU counts and a set of covariates; Tang et al. (2018) and Mao et al. (2020) use the tree decomposition to motivate a divide-and-conquer strategy to increase the statistical power when comparing the OTU composition of groups of samples. In this work, we shall replace the Dirichlet prior in DMM with the DT prior to give a more suitable clustering model for microbiome data. Our main motivation is that when DT is used as the mixing kernel, the model would be more powerful in detecting different clusters. In DMM, if the counts of one OTU are highly variable, the single dispersion parameter would be estimated large in adjustment of this variation. As a result, signals contained in other OTUs would be washed away and the samples would be modeled as drawn from a single cluster. In contrast, the set of dispersion parameters in DT are able to account for different levels of variation across OTUs and thus prevent the signals from being contaminated by the noises.

Specifically, under the sample multinomial sampling scheme as in DMM, let

$$\begin{aligned}
\boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i &\overset{\text{ind}}{\sim} \text{Multi}(N_i, \boldsymbol{p}_i) \\
\boldsymbol{p}_i &\leftarrow \boldsymbol{\theta}_i \\
\boldsymbol{\theta}_i \mid \boldsymbol{\pi}, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^K &\overset{\text{iid}}{\sim} \sum_{k=1}^K \pi_k \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \\
\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim \text{Dir}(\boldsymbol{b}_0),
\end{aligned} \tag{4.8}$$

where

$$(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \overset{\text{iid}}{\sim} \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_0, \boldsymbol{\nu}_0) \times F(\boldsymbol{\tau}), \tag{4.9}$$

$\mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_0, \boldsymbol{\nu}_0)$ the prior for the cluster centroid, $F(\boldsymbol{\tau}) = \prod_{A \in \mathcal{I}} F^A(\tau(A))$ the prior for the within-cluster dispersion. Note that $(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)$ determines the k-th "meta-community". We shall refer to this model as the basic Dirichlet-tree multinomial mixtures (bDTMM).

4.2.3 Discriminative taxa selection

In DMM and bDTMM, all OTUs are treated equally in the clustering procedure. In many applications, however, it is expected that only a (possibly sparse) subset of OTUs determine the underlying clustering. In these cases, it is of scientific interests to identify and report this subset of crucial OTUs along with the clustering profile. Performing OTU selection in clustering is also of statistical importance. DMM and bDTMM introduce cluster-specific parameters on OTUs that are not relevant to the latent class, which may overfit the data and can severely limit their statistical power to identify certain clusters. When applied directly in situations where the signal-to-noise ratio is not very large, DMM and bDTMM tend to identify a few big clusters containing potential sub-clusters.

It is desirable for an OTU selection subroutine in the clustering procedure to have several features. Firstly, since the OTUs are functionally and evolutionarily related, the clustering procedure can be determined by a group of OTUs jointly. As a result, it is desirable for the selection method to provide a shortcut for these groups of OTUs to be selected simultaneously instead of simply treating them as independent. For example, a natural grouping OTUs is provided by the phylogenetic tree—each internal nodes of the tree represents a set of OTUs that are evolutionarily related. Secondly, the selection procedure should be merged with the generative modeling framework to give fast inference.

We next provide an OTU selection method that possesses these features and incorporate it into bDTMM. Specifically, with the tree decomposition, we transform the OTU selection problem into selecting a subset of internal nodes in the phylogenetic tree that are potentially relevant in determining clusters. Formally, for $A \in \mathcal{I}$, let $\gamma(A) \in \{0, 1\}$ be the indicator of whether node A can be contributive to the latent clustering: $\gamma(A) = 1$ if A can play a role in defining clusters, and 0 otherwise. If $\gamma(A) = 1$, we allow different clusters to have cluster-specific branching probabilities at A; otherwise, we "couple" all the clusters at A and force them to share the same branching probability. For this reason, we shall refer to $\gamma(A)$ and $\lambda(A)$ as the "coupling" indicator and the prior "coupling" probability on A. Let $\boldsymbol{\gamma} = \{\gamma(A) : A \in \mathcal{I}\}$ be the collection of coupling indicators of all the internal nodes.

In this section, we consider a generalization to bDTMM: firstly, we incorporate $\boldsymbol{\gamma}$ in the model to allow automatic OTU selection; secondly, we replace the finite mixture in bDTMM with an infinite mixture to avoid presetting the number of clusters. This is equivalent to adopting a Dirichlet process mixture model for the cluster-specific parameters. Let $F(\cdot)$ be a probability measure on $(0, \infty)$, $\delta_x(\cdot)$ the Dirac measure. The model can be written in the following hierarchical form:

- sampling model on $\boldsymbol{y}_i$:

$$\boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i \overset{\text{ind}}{\sim} \text{Multi}(N_i, \boldsymbol{p}_i); \tag{4.10}$$

- priors for the sample-specific probability assignment vector $\boldsymbol{p}_i$:

$$\begin{aligned} \boldsymbol{p}_i &\leftarrow \boldsymbol{\theta}_i \\ \boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i', \boldsymbol{\tau}_i' &\overset{\text{ind}}{\sim} \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i'); \end{aligned} \tag{4.11}$$

- priors for the cluster-specific probability assignment vector:

$$\begin{aligned}(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i') \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim \text{DP}(G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}); \beta);\end{aligned} \tag{4.12}$$

- the base measure in DP:

$$\begin{aligned}G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}) &= \prod_{A \in \mathcal{I}} G_0^A(\theta(A), \tau(A) \mid \gamma(A), \tilde{\theta}(A), \tilde{\tau}(A)) \\ G_0^A(\theta(A), \tau(A) \mid \gamma(A) = 1, \tilde{\theta}(A), \tilde{\tau}(A)) &= \text{Beta}(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A)) \\ &\quad \times F^A(\tau) \\ G_0^A(\theta(A), \tau(A) \mid \gamma(A) = 0, \tilde{\theta}(A), \tilde{\tau}(A)) &= \delta_{(\tilde{\theta}(A), \tilde{\tau}(A))};\end{aligned} \tag{4.13}$$

- priors for the hyperparameters in the base measure: for $A \in \mathcal{I}$,

$$\begin{aligned}\gamma(A) &\overset{\text{ind}}{\sim} \text{Binom}(\lambda(A)) \\ (\tilde{\theta}(A), \tilde{\tau}(A)) &\overset{\text{ind}}{\sim} \text{Beta}(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A)) \times F^A(\tau) \\ \lambda(A) &\overset{\text{ind}}{\sim} \text{Beta}(a_0(A), b_0(A)).\end{aligned} \tag{4.14}$$

We shall refer to this model as the Dirichlet-tree multinomial mixtures (DTMM). The graphical model representation of DTMM is shown in Figure 4.3. Note that G in (4.12) is supported on a countable number of values since samples from a Dirichlet process are discrete, implying ties in the iid samples $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s and thus a clustering on i. This becomes clear with the stick-breaking construction of the Dirichlet process

(Sethuraman, 1994), from which we can rewrite (4.11) and (4.12) as

$$\begin{aligned}
\boldsymbol{p}_i \mid \boldsymbol{\pi}, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^{\infty} &\overset{\text{iid}}{\sim} \sum_{k=1}^{\infty} \pi_k \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \\
\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \text{ where } v_1, v_2, \ldots \mid \beta &\overset{\text{iid}}{\sim} \mathrm{Beta}(1, \beta) \\
(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) &\overset{\text{iid}}{\sim} G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}).
\end{aligned} \tag{4.15}$$

For $i = 1, \ldots, n$, let $c_i \in \mathbb{N}^+$ be the cluster label for the i-th sample such that $\boldsymbol{p}_i \mid c_i, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^{\infty} \sim \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_{c_i}^*, \boldsymbol{\tau}_{c_i}^*)$. We can equivalently illustrate DTMM as in Figure 4.4. For comparison, we can introduce the latent cluster labels to DMM and bDTMM in the same manner and write their graphical model representations as in Figure 4.5 and Figure 4.6. Figure 4.6 and Figure 4.4 illustrate how DTMM generalizes bDTMM.

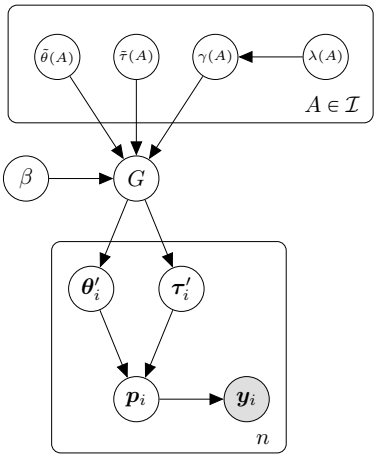

FIGURE 4.3: A graphical model representation of DTMM.

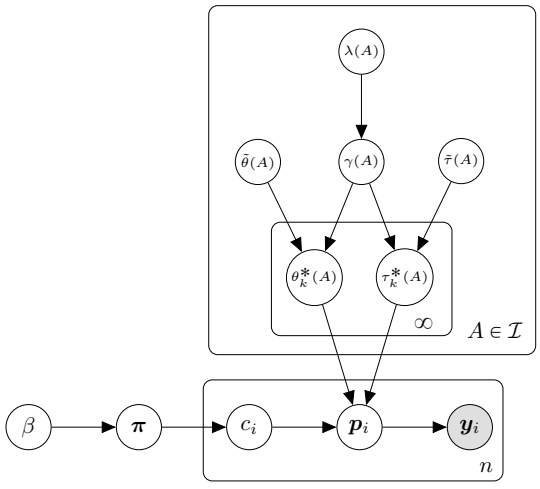

FIGURE 4.4: An alternative graphical model representation of DTMM.

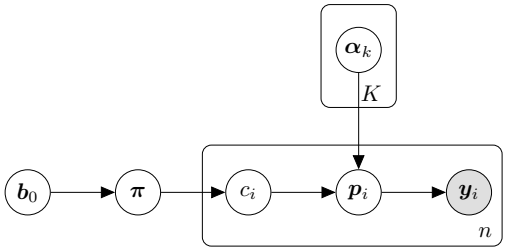

FIGURE 4.5: A graphical representation of DMM.

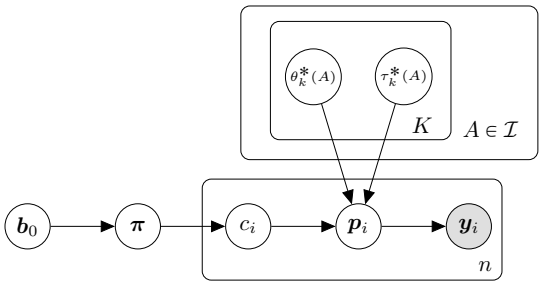

FIGURE 4.6: A graphical model representation of bDTMM.

Prior specification. To complete the model specification, we need to choose $a_0(A)$, $b_0(A)$, $\theta_0(A)$, $\nu_0(A)$ and $F^A(\tau)$ for each $A \in \mathcal{I}$. Ideally, informative prior knowledge shall be incorporated in choosing these parameters. If instead no prior knowledge is available, we treat these parameters (priors) as global such that they do not dependent on A and remove the "(A)'s" from the notations.

For the coupling probability λ, we could set $a_0 = b_0 = 1$ such that λ has a uniform distribution *a priori*, which yields the following prior probability on the $\gamma(A)$'s (Scott and Berger, 2010):

$$\Pr(\boldsymbol{\gamma}) = \frac{1}{M}\binom{M-1}{\sum_{A\in\mathcal{I}}\gamma(A)}^{-1}. \tag{4.16}$$

A default choice for (θ_0, ν_0) is $(0.5, 1)$, which yields the Jeffrey's prior on $\theta_k^*(A)$ and $\theta^*(A)$. For $F(\gamma)$, any prior with a reasonably large support that covers a wide range of dispersion levels can be chosen. For example, we let $F(\tau)$ have density $f(\tau) = (\tau \times 5\log 10)^{-1}\mathbb{1}_{(0.1\leqslant\tau\leqslant 10^4)}$, which is equivalent to putting the Unif$(-1, 4)$ prior on $\log_{10}\tau$. Figure 4.7 shows the histogram of 10^6 draws of τ from this $F(\tau)$. In our software, we use a discrete approximation of this prior induced by drawing $\log_{10}\tau$ uniformly from $\{-1, -0.5, 0, 0.5, 1, \ldots, 4\}$.

Model behavior. In our formulation, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ with a superscript "$*$" are cluster-specific parameters that govern the centroid and the within cluster variance of each

cluster. $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ with a "$\sim$" are parameters that determine the centroid and the variability of the "coupled" base distribution. For $i = 1, \ldots, n$, recall that $c_i \in \mathbb{N}^+$ is the cluster label for the i-th sample. Moreover, let $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$, $\boldsymbol{c}^*$ the set of distinct values in $\boldsymbol{c}$ and $k^* = |\boldsymbol{c}^*|$ the number of distinct clusters. We note that the actual values of c_i bear no significance and thus assume that the c_i's take integer values between 1 and $|\boldsymbol{c}^*|$. At each node $A \in \mathcal{I}$, $\gamma(A)$ serves as a selector: $\tilde{\theta}(A)$ and $\tilde{\tau}(A)$ become relevant only if $\gamma(A) = 0$. If $\gamma(A) = 1$, they are masked and not used by the model. We note that this "masking" happens at the level of the base distribution of the Dirichlet process mixture model. If $\gamma(A) = 0$, the base distribution G_0^A is a point mass. Thus $(\theta_i'(A), \tau_i'(A))$'s must share the same value although $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s may not be the same. In a special case when $\gamma(A) = 0$ for all $A \in \mathcal{I}$, the entire base distribution is a point mass and $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s are all the same. In this case, the cluster labels $\boldsymbol{c}$ are only nominal—the samples are from a single cluster although it is possible that $|\boldsymbol{c}^*| > 1$. Similarly, $\gamma(A)$ as an OTU selector is also nominal—A is not necessarily relevant to clustering even if $\gamma(A) = 1$. In real applications, what we care are not these "nominal" parameters $\boldsymbol{c}$ and $\boldsymbol{\gamma}$ *per se*, but their "actual" counterparts. Specifically, let $g_i \in \mathbb{N}^+$ be the "actual" cluster label of sample i for $i = 1, \ldots, n$, and let $s(A) \in \{0, 1\}$ be the "actual" indicator of whether A is relevant to clustering for $A \in \mathcal{I}$. Moreover, let $\boldsymbol{g} = (g_1, \ldots, g_n)$ and $\boldsymbol{s} = \{s(A) : A \in \mathcal{I}\}$. We have

$$\boldsymbol{g} = \begin{cases} \boldsymbol{c}, & \text{if } \boldsymbol{\gamma} \neq \mathbf{0}_{M-1} \text{ and } \boldsymbol{c} \neq \mathbf{1}_n \\ \mathbf{1}_n, & \text{if } \boldsymbol{\gamma} = \mathbf{0}_{M-1} \text{ or } \boldsymbol{c} = \mathbf{1}_n, \end{cases} \tag{4.17}$$

and

$$\boldsymbol{s} = \begin{cases} \boldsymbol{\gamma}, & \text{if } \boldsymbol{c} \neq \mathbf{1}_n, \text{ and } \boldsymbol{\gamma} \neq \mathbf{0}_{M-1} \\ \mathbf{0}_{M-1}, & \text{if } \boldsymbol{c} = \mathbf{1}_n \text{ or } \boldsymbol{\gamma} = \mathbf{0}_{M-1}. \end{cases} \tag{4.18}$$

Unlike $\boldsymbol{c}$ and $\boldsymbol{\gamma}$, $\boldsymbol{g}$ and $\boldsymbol{s}$ are directly interpretable. For example, $A \in \mathcal{I}$ is relevant

to clustering if and only if $s(A) = 1$. In microbiome applications, it is typically expected that the samples have a latent clustering pattern. Therefore, it is common that $\boldsymbol{g} = \boldsymbol{c}$ and $\boldsymbol{s} = \boldsymbol{\gamma}$.

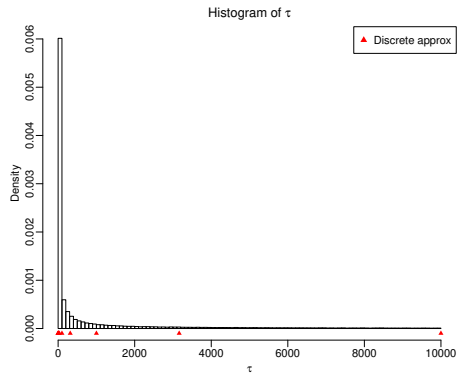

FIGURE 4.7: Histogram of τ. The discrete prior puts equal mass at the red points.

4.2.4 Inference strategy

Under DTMM, we are interested in inferring the nominal cluster labels $\boldsymbol{c}$ and the nominal coupling indicator $\boldsymbol{\gamma}$ from which the actual cluster labels $\boldsymbol{g}$ and the actual coupling indicators $\boldsymbol{s}$ can be obtained. Let $\boldsymbol{y}_{-i}$ denote all the observations other than $\boldsymbol{y}_i$. Bayesian inference for DTMM can be achieved by constructing a Markov chain that converges to the joint posterior of $(\boldsymbol{c}, \boldsymbol{\gamma})$. We shall see that techniques designed for Dirichlet process mixture models in general, such as those described in Neal (2000) or Ishwaran and James (2001), can be applied here.

For $c \in \boldsymbol{c}^*$, let $\boldsymbol{\psi}_c^* = (\boldsymbol{\theta}_c^*, \boldsymbol{\tau}_c^*)$ be the parameters that define the cluster indicated by c (we also let $\boldsymbol{\psi}_c^*(A) = (\theta_c^*(A), \tau_c^*(A))$ for $A \in \mathcal{I}$). Similarly, let $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}})$ be the shared parameters at the coupled nodes and $\tilde{\boldsymbol{\psi}}(A) = (\tilde{\theta}(A), \tilde{\tau}(A))$ for $A \in \mathcal{I}$. The set of unknown parameters in DTMM is $\{\{\boldsymbol{\theta}_i, c_i\}_{i=1}^n, \{\boldsymbol{\psi}_c^*\}_{c=1}^{k^*}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}, \beta, \lambda\}$. In this work, we construct a collapsed Gibbs sampler that iteratively samples from the joint posterior of $(\boldsymbol{c}, \boldsymbol{\gamma}, \beta, \lambda)$. The key to our inference strategy is to compute the marginal likelihoods of samples from a given cluster, integrating out both the sample-specific

parameter $\boldsymbol{\theta}_i$ and the cluster-specific parameter $\boldsymbol{\psi}_c^*$. This can be achieved numerically due to two facts. Firstly, the beta-binomial conjugacy makes it easy to integrate out the sample-specific compositional probabilities $\boldsymbol{\theta}_i$. Secondly, the tree-based decomposition of the Dirichlet distribution and the multinomial likelihood provides a divide-and-conquer strategy to marginalize out the high-dimensional cluster-specific parameters $\boldsymbol{\psi}_c^*$ through performing a series of low-dimensional integrals at the internal nodes of the tree.

Specifically, for any $c \in \boldsymbol{c}^*$, let $\boldsymbol{Y}_c^I = \{\boldsymbol{y}_i : c_i = c, i \in I\}$ be a set of samples in cluster c where $I \subset [n] := \{1, \ldots, n\}$. We also let $\boldsymbol{Y}_c = \boldsymbol{Y}_c^{[n]}$ be the set of all samples in cluster c and $\boldsymbol{Y}_c^{-i} = \boldsymbol{Y}_c^{[n]\setminus\{i\}}$ be the set of samples in cluster c excluding sample i. For $A \in \mathcal{I}$, let $\mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A), \tilde{\boldsymbol{\psi}}(A))$ be the marginal likelihood of $\boldsymbol{Y}_c^I$ at node A by marginalizing out the sample-specific parameters. The beta-binomial conjugacy yields

$$\begin{aligned}
&\mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A), \tilde{\boldsymbol{\psi}}(A)) \\
&= \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta_c^*(A)\tau_c^*(A) + y_i(A_l), (1-\theta_c^*(A))\tau_c^*(A) + y_i(A_r))}{B(\theta_c^*(A)\tau_c^*(A), (1-\theta_c^*(A))\tau_c^*(A))}.
\end{aligned} \tag{4.19}$$

We then further integrate out $\boldsymbol{\psi}_c^*(A)$ to obtain the marginal likelihood of $\boldsymbol{Y}_c^I$ at node A given only the coupling indicators and the base parameters:

$$\begin{aligned}
\mathcal{L}_1^A(\boldsymbol{Y}_c^I) &:= \iint \mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)) d\Pi(\boldsymbol{\psi}_c^*(A) \mid \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)) \\
&= \iint \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta(A)\tau(A) + y_i(A_l), (1-\theta(A))\tau(A) + y_i(A_r))}{B(\theta(A)\tau(A), (1-\theta(A))\tau(A))} \\
&\qquad \times \frac{\theta(A)^{\theta_0(A)\nu_0(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu_0(A)-1}}{B(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A))} d\theta(A) dF^A(\tau),
\end{aligned} \tag{4.20}$$

$$\mathcal{L}_0^A(\boldsymbol{Y}_c^I \mid \tilde{\boldsymbol{\psi}}(A)) := \iint \mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A)=0, \tilde{\boldsymbol{\psi}}(A)) d\Pi(\boldsymbol{\psi}_c^*(A) \mid \gamma(A)=0, \tilde{\boldsymbol{\psi}}(A))$$

$$= \prod_{\{i \in I: c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\tilde{\theta}(A)\tilde{\tau}(A) + y_i(A_l), (1-\tilde{\theta}(A))\tilde{\tau}(A) + y_i(A_r))}{B(\tilde{\theta}(A)\tilde{\tau}(A), (1-\tilde{\theta}(A))\tilde{\tau}(A))}. \tag{4.21}$$

Note that the integrals in (B.1) are two-dimensional integrals that are easy to evaluate numerically. In comparison, to perform a fully Bayesian inference for DMM, the high dimensional cluster centroids $\boldsymbol{\alpha}_k$'s in (4.2) has to either be integrated out directly or incorporated into the MCMC procedure. With these marginal likelihoods, we can summarize the full conditionals of the parameters as follows. Details on the derivation of these full conditionals are given in the online supplementary materials A.

The full conditional of $\boldsymbol{\gamma}$: For each $A \in \mathcal{I}$, the Bayes factor comparing $\gamma(A) = 1$ versus $\gamma(A) = 0$ given $\boldsymbol{c}$ can be written as

$$M_{10}(A \mid \boldsymbol{\gamma}^{-A}, \boldsymbol{c}, \beta, \lambda) = M_{10}(A \mid \boldsymbol{c}) = \frac{\prod_{c \in \boldsymbol{c}^*} \mathcal{L}_1^A(\boldsymbol{Y}_c)}{\int \mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A)) d\Pi(\tilde{\boldsymbol{\psi}}(A))}. \tag{4.22}$$

It follows that

$$\Pr(\gamma(A) = 1 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{-A}, \boldsymbol{c}, \beta, \lambda) = \frac{\lambda M_{10}(A \mid \boldsymbol{c})}{(1-\lambda) + \lambda M_{10}(A \mid \boldsymbol{c})}. \tag{4.23}$$

The full conditional of $\boldsymbol{c}$: For $i = 1, \ldots, n$, let $\boldsymbol{c}_{-i} = \boldsymbol{c} \backslash \{c_i\}$. Following the discussion in Neal (2000), we can write the prior conditional distribution of c_i given

$\boldsymbol{c}_{-i}$ as

$$\begin{aligned}\Pr(c_i = c \text{ for some } c \in \boldsymbol{c}_{-i} \mid \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) &= \frac{n_{-i,c}}{n-1+\beta} \\ \Pr(c_i \neq c_j \text{ for all } j \neq i \mid \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) &= \frac{\beta}{n-1+\beta},\end{aligned} \tag{4.24}$$

where $n_{-i,c}$ represents the number of samples in the cluster with label c excluding sample i. After conditioning on the data, these probabilities become

$$\begin{aligned}\Pr(c_i = c \text{ for some } c \in c_{-i} \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) &\propto n_{-i,c} \times \frac{\mathcal{L}_1(\boldsymbol{y}_i, \boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})}{\mathcal{L}_1(\boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})} \\ \Pr(c_i \neq c_j \text{ for all } j \neq i \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) &\propto \beta \times \mathcal{L}_1(\boldsymbol{y}_i \mid \boldsymbol{\gamma}),\end{aligned} \tag{4.25}$$

where for any $\boldsymbol{y}_i, \ldots, \boldsymbol{y}_l$ in the same cluster,

$$\mathcal{L}_1(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l \mid \boldsymbol{\gamma}) = \prod_{\{A \in \mathcal{I}: \gamma(A)=1\}} \mathcal{L}_1^A(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l). \tag{4.26}$$

If $\gamma(A) = 0$ for every $A \in \mathcal{I}$, we let $\mathcal{L}_1(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l) = 1$. Note that given the coupling status $\boldsymbol{\gamma}$, the posterior conditional distribution of c_i only depends on the likelihoods of the data at nodes with $\gamma(A) = 1$. Therefore, to update the cluster label for any observation $\boldsymbol{y}_i$, we only need to focus on the nodes with $\gamma(A) = 1$ and at each of these nodes compute: (i) the marginal likelihood of $\boldsymbol{y}_i$ and (ii) for each $c \in \boldsymbol{c}_{-i}$, the conditional likelihood of $\boldsymbol{y}_i$ given $\boldsymbol{Y}_c^{-i}$. All nodes with $\gamma(A) = 0$ and thus the parameters at these nodes are essentially "nuisance" for the cluster labels.

The full conditional of β: Instead of fixing β, one can put a prior on it and incorporate it into the Gibbs sampler. For example, when gamma priors are used, Escobar and West (1995) update β using a data augmentation trick. When arbitrary priors are used, β can be updated by reparameterizing $b = \frac{\beta}{\beta+1}$ (Hoff et al., 2006). Specifically, let $\pi(b)$ be the induced prior on b, we have

$$\pi(b \mid \boldsymbol{Y}, \boldsymbol{c}, \boldsymbol{\gamma}, \lambda) \propto \pi(b) \times \left(\frac{b}{1-b}\right)^{|\boldsymbol{c}^*|} \frac{\Gamma(b/(1-b))}{\Gamma(b/(1-b)+n)}. \tag{4.27}$$

The full conditional of λ: By the beta-binomial conjugacy,

$$\lambda \mid \boldsymbol{Y}, \boldsymbol{c}, \boldsymbol{\gamma}, \beta \sim \text{Beta}\left(a_0 + \sum_{A \in \mathcal{I}} \gamma(A), b_0 + \sum_{A \in \mathcal{I}} (1 - \gamma(A))\right). \tag{4.28}$$

Details for implementing the Gibbs sampler are summarized in Algorithm 2. After running the chain for T iterations, we discard the first B samples as burn-in and obtain $(T - B)$ posterior samples denoted as $\big[\{\boldsymbol{c}^{(B+1)}, \boldsymbol{\gamma}^{(B+1)}, \beta^{(B+1)}, \lambda^{(B+1)}\}, \ldots, \{\boldsymbol{c}^{(T)}, \boldsymbol{\gamma}^{(T)}, \beta^{(T)}, \lambda^{(T)}\}\big]$. Based on these posterior samples, we can compute the posterior samples for $\boldsymbol{g}$ and $\boldsymbol{s}$ based on (4.17) and (4.18). We denote these posterior samples as $[\{\boldsymbol{g}^{(B+1)}, \boldsymbol{s}^{(B+1)}\}, \ldots, \{\boldsymbol{g}^{(T)}, \boldsymbol{s}^{(T)}\}]$

For each sample $\boldsymbol{g}^{(t)}$, let $\Gamma^{(t)}$ be the corresponding $n \times n$ association matrix whose (i_1, i_2) element is 1 if $g_{i_1}^{(t)} = g_{i_2}^{(t)}$ and 0 otherwise. Element-wise average of $\Gamma^{(B+1)}, \ldots, \Gamma^{(T)}$ provides an estimation $\hat{\Pi}$ of the pairwise clustering probability matrix Π whose (i_1, i_2) element is $\Pr(\boldsymbol{y}_{i_1}$ and $\boldsymbol{y}_{i_2}$ in the same cluster). To yield a representative clustering, we can report the least-squares model-based clustering (Dahl, 2006), defined as

$$\boldsymbol{C}_{LS} = \underset{\{\boldsymbol{g}^{(t)}: B < t \leqslant T\}}{\arg\min} \sum_{1 \leqslant i_1 \leqslant n} \sum_{1 \leqslant i_2 \leqslant n} (\Gamma_{i_1 i_2}^{(t)} - \hat{\Pi}_{i_1 i_2})^2. \tag{4.29}$$

$\boldsymbol{C}_{LS}$ has the advantage that it incorporates information from all posterior samples while output one of the observed clustering in the Markov Chain (Dahl, 2006). Other representative clusterings such as the MAP clustering or the clustering given by the last iteration are also frequently used.

We can also portray the cluster centroids given any representative clustering and the corresponding coupling indicators. Suppose that $\boldsymbol{c}_{rep} = \boldsymbol{g}^{(t_0)}$ for $B < t_0 \leqslant T$ is one of the representative clusterings and let $\boldsymbol{\gamma}_{rep} = \boldsymbol{s}^{(t_0)}$. For the k-th resulting cluster defined by $\boldsymbol{c}_{rep}$, $1 \leqslant k \leqslant |\boldsymbol{c}_{rep}^*|$, the posterior mean of the branching probability at

Algorithm 2 A collapsed Gibbs sampler for DTMM

procedure GIBBS($B, T, \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$) ▷ B : burn-in; T: total number of iterations.

Initialize $\boldsymbol{c}^{(0)}, \boldsymbol{\gamma}^{(0)}, \beta^{(0)}, \gamma^{(0)}$.

($\boldsymbol{\gamma}^{(0)}, \beta^{(0)}, \gamma^{(0)}$ can be initialized with random draws from their prior, $\boldsymbol{c}^{(0)}$ can be initialized with the clustering result of some distance-based algorithms such as PAM.)

for $t = 1, 2, \ldots, T$ **do**

[1] Update the coupling indicators:

for $A \in \mathcal{I}$ **do**

Compute $M_{10}^{(t-1)}(A \mid \boldsymbol{c}^{(t-1)})$ as defined by (4.23).

Draw a new value for $\gamma^{(t)}(A)$:

$$\gamma^{(t)}(A) \sim \text{Binom}\left(1, \frac{\lambda^{(t-1)} M_{10}^{(t-1)}(A \mid \boldsymbol{c}^{(t-1)})}{(1-\lambda^{(t-1)}) + \lambda^{(t-1)} M_{10}^{(t-1)}(A \mid \boldsymbol{c}^{(t-1)})}\right).$$

end for

[2] Update the cluster labels:

for $i = 1, 2, \ldots, n$ **do**

Draw a new value for $c_i^{(t)}$ from

$$c_i \mid c_1^{(t)}, \ldots, c_{i-1}^{(t)}, c_{i+1}^{(t-1)}, \ldots, c_n^{(t-1)}, \boldsymbol{Y}, \boldsymbol{\gamma}^{(t)}, \beta^{(t-1)}$$

as defined by (4.25).

end for

[3] Update the Dirichlet process precision parameter:

Draw value b^{new} from $b \sim \pi(b \mid \boldsymbol{Y}, \boldsymbol{c}^{(t)}, \boldsymbol{\gamma}^{(t)}, \lambda^{(t-1)})$ as defined in (4.27). Let $\beta^{(t)} = \frac{b^{\text{new}}}{1-b^{\text{new}}}$.

[4] Update the prior coupling probabilities:

Draw $\lambda^{(t)}$ from $\lambda \mid \boldsymbol{Y}, \boldsymbol{c}^{(t)}, \boldsymbol{\gamma}^{(t)}, \beta^{(t)}$ as defined in (4.28).

end for

return $\left[\{\boldsymbol{c}^{(B+1)}, \boldsymbol{\gamma}^{(B+1)}, \beta^{(B+1)}, \lambda^{(B+1)}\}, \ldots, \{\boldsymbol{c}^{(T)}, \boldsymbol{\gamma}^{(T)}, \beta^{(T)}, \lambda^{(T)}\}\right]$.

end procedure

$A \in \mathcal{I}$ given the coupling status $\gamma_{rep}(A)$ can be written as

$$\begin{aligned}
\mathbb{E}[\theta_k^*(A) \mid \boldsymbol{Y}, \gamma_{rep}(A) = 1] &= \iint \theta_k^*(A) \times \pi(\theta_k^*(A), \tau_k^*(A) \mid \boldsymbol{Y}_k) d\theta_k^*(A) d\tau_k^*(A) \\
\mathbb{E}[\tilde{\theta}(A) \mid \boldsymbol{Y}, \gamma_{rep}(A) = 0] &= \iint \tilde{\theta}(A) \times \pi(\tilde{\theta}(A), \tilde{\tau}(A) \mid \boldsymbol{Y}) d\tilde{\theta}(A) d\tilde{\tau}(A)
\end{aligned} \tag{4.30}$$

where

$$\begin{aligned}
\pi(\theta_k^*(A), \tau_k^*(A) \mid \boldsymbol{Y}_k) \propto & \prod_{\boldsymbol{y}_i \in \boldsymbol{Y}_K} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta_k^*(A)\tau_k^*(A) + y_i(A_l), (1-\theta_k^*(A))\tau_k^*(A) + y_i(A_r))}{B(\theta_k^*(A)\tau_k^*(A), (1-\theta_k^*(A))\tau_k^*(A))} \\
& \times \pi(\theta_k^*(A), \tau_k^*(A)) \\
\pi(\tilde{\theta}(A), \tilde{\tau}(A) \mid \boldsymbol{Y}) \propto & \mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A)) \times \pi(\tilde{\theta}(A), \tilde{\tau}(A)).
\end{aligned} \tag{4.31}$$

Note that (4.30) involves two-dimensional integrals that can be numerically approximated accurately. Let $\bar{\boldsymbol{p}}_k = (\bar{p}_{k1}, \ldots, \bar{p}_{kM})$ be the posterior mean of $\boldsymbol{p}_k^*$, which is the centroid of the k-th cluster. For $\omega_j \in \Omega$, let $A_0^j \to A_1^j \to \cdots \to A_{l_j}^j \to \omega_j$ be the unique path in $\mathcal{T}$ connecting Ω and ω_j. Then

$$\begin{aligned}
\bar{p}_{kj} = \prod_{l=0}^{l_j} & \mathbb{E}[\theta_k^*(A_l^j) \mid \boldsymbol{Y}, \gamma_{rep}(A_l^j) = 1]^{\mathbb{1}(\gamma_{rep}(A_l^j)=1)} \\
& \times \mathbb{E}[\tilde{\theta}(A_l^j) \mid \boldsymbol{Y}, \gamma_{rep}(A_l^j) = 0]^{\mathbb{1}(\gamma_{rep}(A_l^j)=0)}.
\end{aligned} \tag{4.32}$$

Note that $\bar{\boldsymbol{p}}_k$ only characterizes the centroid of the k-th cluster. To characterize the within-cluster dispersion, we need to look at the posterior distribution of $\tau_k^*(A)$ or $\tilde{\tau}(A)$ for $A \in \mathcal{I}$, which is available through marginalizing out $\theta_k^*(A)$ or $\tilde{\theta}(A)$ in (4.31).

4.2.5 Sample classification for microbiome data with DTMM

The DTMM framework can also be used in the supervised setting to achieve sample classification based on a training microbiome dataset. Without loss of generality, suppose that the training dataset contains microbiome samples from K classes:

$\{(\boldsymbol{y}_1, c_1), (\boldsymbol{y}_2, c_2), \ldots, (\boldsymbol{y}_n, c_n)\}$, $1 \leqslant c_i \leqslant K$. We consider the following generative model of $\boldsymbol{Y}$:

$$\begin{aligned} \boldsymbol{y}_i \mid c_i = k &\overset{\text{ind}}{\sim} \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \\ (\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) &\overset{\text{iid}}{\sim} G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}) \\ \Pr(c_i = k) &= \pi_k, \end{aligned} \tag{4.33}$$

where G_0 and the hyperparameters of G_0 are specified as in (4.13) and (4.14).

Let $\boldsymbol{y}^{new}$ be a new microbiome sample from (4.33). It follows that

$$\begin{aligned} &\Pr(c^{new} = k \mid \boldsymbol{y}^{new}, \boldsymbol{Y}) \\ \propto\ &\pi_k \cdot \mathcal{L}(\boldsymbol{y}^{new} \mid c^{new} = k, \boldsymbol{Y}) \\ \propto\ &\sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \iint \mathcal{L}(\boldsymbol{y}^{new} \mid \boldsymbol{\psi}_k^*, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}, \boldsymbol{Y}_k) d\Pi(\boldsymbol{\psi}_k^* \mid \boldsymbol{Y}_k, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}) d\Pi(\tilde{\boldsymbol{\psi}} \mid \boldsymbol{Y}). \end{aligned} \tag{4.34}$$

Note that (4.34) can be numerically evaluated in a way similar to the marginal likelihood evaluation in DTMM. Details of doing inference for classification under the DTMM framework can be found in the online supplementary materials A.3.

4.3 Numerical examples

4.3.1 Simulation studies

Simulation setup

In this section, we carry out a series of numerical studies to evaluate the performance of DTMM and compare it to several other methods for clustering microbiome count data—namely, the Dirichlet multinomial mixtures (DMM) (Holmes et al., 2012), the k-means algorithm (K-ms) (Lloyd, 1982), the partitioning around medoids algorithm (PAM) (Kaufman and Rousseeuw, 2009), hierarchical clustering (Hclust) (Kaufman and Rousseeuw, 2009), and spectral clustering (Spec) (Ng et al., 2002).

In the numerical examples, we simulate datasets with n samples and six OTUs. In each dataset, the n samples are denoted as $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})$, $i = 1, \ldots, n$, which

are generated from the following model:

$$\begin{aligned} \boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i &\overset{\text{ind}}{\sim} \text{Multi}(N_i, \boldsymbol{p}_i) \\ \boldsymbol{p}_i &\overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k \cdot H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) \\ N_i &\overset{\text{iid}}{\sim} \text{Neg-Binom}(m, s), \end{aligned} \tag{4.35}$$

where the mixture kernel $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k)$ is a distribution on the 5-simplex with parameter $\boldsymbol{\beta}_k$. We take the tree in Figure 4.8 as the "phylogenetic tree" over the six OTUs and use it as the $\mathcal{T}$ when fitting DTMM. We consider 5 different simulation scenarios by choosing different mixture kernels $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k)$ in (4.35). In each scenario, we let $n = 90$ or 180, $K = 3$ and $(\pi_1, \pi_2, \pi_3) = (\frac{4}{9}, \frac{3}{9}, \frac{2}{9})$. Parameters for the negative-binomial distribution are chosen as $m = 15000$, $s = 20$ such that the generated total counts has mean 15000 and standard deviation 3346, with 95% of them fall into the range $(9158, 22258)$. In the 5 simulation scenarios, the mixture kernels are chosen as follows:

I. **Dirichlet-tree kernel.** We first fix $\mathcal{T} = \mathcal{T}_6$ as in Figure 4.8 and let $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k, \boldsymbol{\tau}_k)$ such that DTMM is the "true" model. The parameters $(\boldsymbol{\theta}_k, \boldsymbol{\tau}_k)$ are chosen such that the branching probabilities at the 5 internal nodes have the Beta distributions as shown in Figure 4.9, where $\boldsymbol{\nu}_1 = (10\alpha, 2\alpha)$, $\boldsymbol{\nu}_2 = (6\alpha, 6\alpha)$, $\boldsymbol{\nu}_3 = (2\alpha, 10\alpha)$ and $\gamma = 0.1$. We write this specific family of Dirichlet-tree distributions as $\text{DT}_{\mathcal{T}_6}(\boldsymbol{\nu}_k; \alpha; \tau)$, $k = 1, 2, 3$. Note that when $\alpha = \gamma$, the Dirichlet-tree distribution becomes the Dirichlet distribution. In this case, only the branching probabilities at node C contribute to the clustering. The signals are thus local to a single node. Three signal levels are considered by letting $\alpha = 1, 3, 6$.

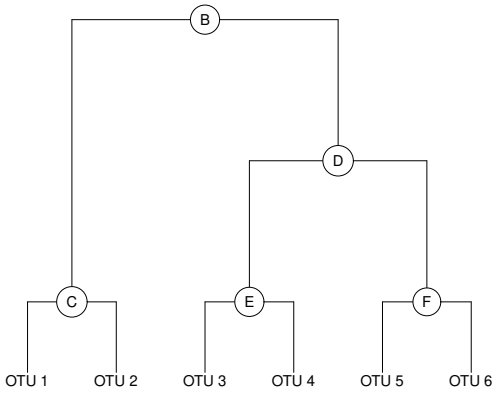

FIGURE 4.8: The phylogenetic tree for the simulated examples.

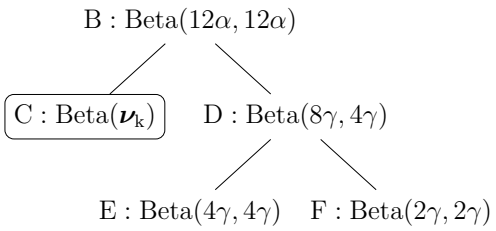

FIGURE 4.9: Distributions for the branching probabilities in Scenario I.

II. **Dirichlet kernel.** In this scenario, we let $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{Dir}(\boldsymbol{\alpha}_k)$ such that DMM is the "true" model that generates the data. In this case, DTMM is still correct but is over-specified. We let $\boldsymbol{\alpha}_1 = (2, 2, 5, 2, 3, 1) \cdot \alpha_0$, $\boldsymbol{\alpha}_2 = (2, 4, 3, 2, 1, 3) \cdot \alpha_0$ and $\boldsymbol{\alpha}_3 = (2, 6, 1, 2, 2, 2) \cdot \alpha_0$ for $\alpha_0 > 0$ such that all six OTUs are active in differentiating the clusters and the signals are global. We consider three signal levels with $\alpha_0 = 1, 3, 5$.

In the following three examples, we evaluate the performance of DTMM when the model is misspecified. Let $\boldsymbol{p} = (p_1, \ldots, p_6) \in \mathbb{S}^5$. We say that $\boldsymbol{p}$ has the logistic normal distribution (Atchison and Shen, 1980) and denote as $\boldsymbol{p} \sim \text{Logit-Norm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if

$$\boldsymbol{x} = \left(\log\left(\frac{p_1}{p_6}\right), \ldots, \log\left(\frac{p_5}{p_6}\right) \right)^\top$$

$$\boldsymbol{x} \overset{\text{ind}}{\sim} \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

In Scenarios III, IV and V, we let $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{Logit-Norm}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. In these examples, we assume that the phylogenetic tree $\mathcal{T}_6$ provides some insights on the covariate structures of the OTUs.

III. **Logistic-normal approximations to the Dirichlet-tree kernel.** Consider

the Dirichlet-tree kernels $q_k = \mathrm{DT}_{\mathcal{T}_6}(\boldsymbol{\nu}_k; \alpha; \tau)$ as in Scenario I with $\gamma = 0.5$ for $k = 1, 2, 3$. In this example, we let

$$H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \arg\min_{h \in \mathbb{L}^5} D_{\mathrm{KL}}(q_k \parallel h)$$

where $D_{\mathrm{KL}}(q_k \parallel h)$ is the Kullback-Leibler divergence from h to q_k, $\mathbb{L}^5$ the set of logistic-normal distributions on $\mathbb{S}^5$. It is shown in the online supplementary materials A.1. that $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{Logit-Norm}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where

$$\boldsymbol{\mu}_k = \mathbb{E}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right], \quad \boldsymbol{\Sigma}_k = \mathbb{V}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right],$$

which are available in closed form by the properties of the exponential family. In this scenario, DTMM is not the correct model, but we expect that it is not severely misspecified. Three signal levels are considered with $\alpha = 3, 6, 9$.

IV. **Logistic-normal kernel (single node).** In this scenario, we let $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{Logit-Norm}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Sigma}_k = \mathrm{diag}(0.05, 0.05, 1, 1, 1)$, $\boldsymbol{\mu}_1 = (3, 1, a, b, 0)$, $\boldsymbol{\mu}_2 = (2.43, 2.43, a, b, 0)$ and $\boldsymbol{\mu}_3 = (1, 3, a, b, 0)$. The $\boldsymbol{\mu}_k$'s are chosen such that only node C in $\mathcal{T}_6$ is relevant for clustering. The diagonal covariance matrix suggests that DTMM is misspecified (in comparison, in III the covariance matrices are dense). We consider three signal levels with $(a, b) = (5, 3), (2, 2)$ and $(1, 1)$. Note that when the relative abundance of OTU_3 and OTU_4 are high, the compositional nature of the data implies that fewer counts are generated for OTU_1 and OTU_2 that determine the clustering, resulting in a low signal-to-noise ratio.

V. **Logistic-normal kernel (multiple nodes).** Similar to scenario IV, we let $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k) = \text{Logit-Norm}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Sigma}_k = \mathrm{diag}(1, 1, 0.05, 0.05, 0.05)$, $\boldsymbol{\mu}_1 = (c, d, 3.5, 3, 2.5)$, $\boldsymbol{\mu}_2 = (c, d, 2.5, 3.5, 3)$ and $\boldsymbol{\mu}_3 = (c, d, 3, 2.5, 3.5)$. In this

case, the $\boldsymbol{\mu}_k$'s are chosen such that the clusters are determined by the relative abundance of OTU_3, OTU_4 and OTU_5, which are reflected in nodes D, E and F in $\mathcal{T}_6$ in Figure 4.8. Three signal levels with $(c, d) = (6, 6), (3, 3)$ and $(1, 1)$ are considered.

Table 4.2: Mixture kernels for generating the simulated datasets.

	Kernel	Signal		$\boldsymbol{\beta}_k$	
		Level	Parameter		
I	DT	W	$\alpha = 1$	$(12\alpha, 12\alpha)$; $\boldsymbol{\nu}_1 = (10\alpha, 2\alpha)$, $\boldsymbol{\nu}_2 = (6\alpha, 6\alpha)$, $\boldsymbol{\nu}_3 = (2\alpha, 10\alpha)$; $(8\gamma, 4\gamma)$; $(4\gamma, 4\gamma)$; $(2\gamma, 2\gamma)$	$\gamma = 0.1$
		M	$\alpha = 3$		
		S	$\alpha = 6$		
II	Dir	W	$\alpha_0 = 1$	$\boldsymbol{\alpha}_1 = (2, 2, 5, 2, 3, 1) \cdot \alpha_0$	
		M	$\alpha_0 = 3$	$\boldsymbol{\alpha}_2 = (2, 4, 3, 2, 1, 3) \cdot \alpha_0$	
		S	$\alpha_0 = 6$	$\boldsymbol{\alpha}_3 = (2, 6, 1, 2, 2, 2) \cdot \alpha_0$	
III	LN	W	$\alpha = 3$	$q_k = \text{DT}_{\mathcal{T}_6}(\boldsymbol{\nu}_k; \alpha; 0.5)$ $\boldsymbol{\nu}_1 = (10\alpha, 2\alpha)$ $\boldsymbol{\nu}_2 = (6\alpha, 6\alpha)$ $\boldsymbol{\nu}_3 = (2\alpha, 10\alpha)$	$\boldsymbol{\mu}_k = \mathbb{E}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right]$ $\boldsymbol{\Sigma}_k = \mathbb{V}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right]$
		M	$\alpha = 6$		
		S	$\alpha = 9$		
IV	LN	W	$a = 5, b = 3$	$\boldsymbol{\mu}_1 = (3, 1, a, b, 0)$	$\Sigma_{1,2,3} = \begin{pmatrix} 0.05 & & & & \\ & 0.05 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix}$
		M	$a = 2, b = 2$	$\boldsymbol{\mu}_2 = (2.43, 2.43, a, b, 0)$,	
		S	$a = 1, b = 1$	$\boldsymbol{\mu}_3 = (1, 3, a, b, 0)$	
V	LN	W	$c = 6, d = 6$	$\boldsymbol{\mu}_1 = (c, d, 3.5, 3, 2.5)$	$\Sigma_{1,2,3} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 0.05 & & \\ & & & 0.05 & \\ & & & & 0.05 \end{pmatrix}$
		M	$c = 3, d = 3$	$\boldsymbol{\mu}_2 = (c, d, 2.5, 3.5, 3)$,	
		S	$c = 1, d = 1$	$\boldsymbol{\mu}_3 = (c, d, 3, 2.5, 3.5)$	

Mixture kernels for generating datasets in the 5 simulation scenarios are summarized in Table 4.2. In each scenario, a "null" case is also considered by setting $K = 1$ in the case with the medium signal level. For each (kernel, signal level) combination, we conduct 100 rounds of simulations. For each simulated dataset with $K = 3$, we calculate the following R^2 as a measure of the strength of the signal (Anderson,

2001):

$$R^2 = \frac{\text{SSW}}{\text{SST}} = \frac{\sum_{k=1}^{3}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{\text{BC}}(\boldsymbol{y}_i, \boldsymbol{y}_j)^2 \epsilon_{ij}^k / n_k}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{\text{BC}}(\boldsymbol{y}_i, \boldsymbol{y}_j)^2 / n},$$

where $d_{\text{BC}}(\cdot, \cdot)$ is the Bray-Curtis dissimilarity, n_k the number of samples in cluster k, $\epsilon_{ij}^k = 1$ if the samples i and j are both in cluster k and 0 otherwise. The average R^2's of the 100 simulated datasets in each experiment are reported in Table 4.3 and Table 4.4.

In each simulation round, we run the Gibbs sampler for DTMM for 2000 iterations and discard the first half of the chain as burn-in. The priors and hyper-parameters for DTMM are set to the recommended choice in Section 4.2.3. The initial values for the clustering labels in the Markov chain are set to the labels of running the k-means algorithm with $k = 5$. For DTMM, we output $\boldsymbol{c}_{LS}$ as a representative clustering. For PAM and Hclust, we use the Bray-Curtis dissimilarity on the relative abundance as the underlying distance measure between samples. For all competitors other than DMM, the number of clusters is required as a tuning parameter, we set this parameter to the truth 3 when running these methods.

Analyses

To compare the performance of different methods, we compute the Jaccard index (Jaccard, 1912) between the clusters obtained by each method and the truth. For a specific clustering $\boldsymbol{c}$ and the true clustering $\boldsymbol{c}_0$, the Jaccard index between $\boldsymbol{c}$ and $\boldsymbol{c}_0$ is defined as $J(\boldsymbol{c}, \boldsymbol{c}_0) = \mathcal{N}_{\boldsymbol{c}\cap\boldsymbol{c}_0}/\mathcal{N}_{\boldsymbol{c}\cup\boldsymbol{c}_0}$, where $\mathcal{N}_{\boldsymbol{c}\cap\boldsymbol{c}_0}$ is the number of pairs of samples that are in the same cluster under both $\boldsymbol{c}$ and $\boldsymbol{c}_0$, $\mathcal{N}_{\boldsymbol{c}\cup\boldsymbol{c}_0}$ the number of pairs of samples that are in the same cluster under at least one of $\boldsymbol{c}$ and $\boldsymbol{c}_0$. When $\boldsymbol{c}$ gives the same clustering as $\boldsymbol{c}_0$, $J(\boldsymbol{c}, \boldsymbol{c}_0) = 1$. In each simulation scenario, we compare the root

mean squared error of each method m: $\text{RMSE}^{(m)} = \sqrt{\sum_{r=1}^{100}[J(\boldsymbol{c}_r^{(m)}, \boldsymbol{c}_0) - 1]^2/100}$, where $\boldsymbol{c}_r^{(m)}$ is the clustering obtained by method m in simulation round r. As some references, let $\boldsymbol{c}_0 = (1 \cdot \mathbf{1}_{40}^\top, 2 \cdot \mathbf{1}_{30}^\top, 3 \cdot \mathbf{1}_{20}^\top)$, $\boldsymbol{c}_1 = (1 \cdot \mathbf{1}_{90}^\top)$, $\boldsymbol{c}_2 = (1 \cdot \mathbf{1}_{30}^\top, 2 \cdot \mathbf{1}_{30}, 3 \cdot \mathbf{1}_{30})$ and $\boldsymbol{c}_3 = (1 \cdot \mathbf{1}_{40}^\top, 2 \cdot \mathbf{1}_{50})$, where $\mathbf{1}_n$ is the n-dimensional vector with all element equal to 1. We have $\sqrt{[J(\boldsymbol{c}_1, \boldsymbol{c}_0) - 1]^2} = 0.65$, $\sqrt{[J(\boldsymbol{c}_2, \boldsymbol{c}_0) - 1]^2} = 0.50$ and $\sqrt{[J(\boldsymbol{c}_3, \boldsymbol{c}_0) - 1]^2} = 0.30$. Table 4.3 and Table 4.4 shows the RMSE of DTMM and the competitors under all simulation scenarios. Boxplots of the Jaccard index reported by each method are shown in Figure B.1 and Figure B.2 in the online supplementary materials A.2.

When $K = 3$, DTMM is always one of the top two methods under comparison. When it's not the best method, its performance is often close to the best. Without utilizing the information provided by the phylogenetic tree, all competitors of DTMM provide limited insights on the clustering when the signal is weak or medium. Moreover, these competitors rely on global distance measures between samples and treat the six OTUs equivalently. As a result, in scenario like I and IV where the signal is local to a single internal node of the phylogenetic tree, these methods have poor performance. Even in scenario V where half of the OTUs are relevant for clustering, these methods still suffer unless the signal is very strong. In scenario II where the signal is global, all methods perform reasonably well. Note that in this scenario, DTMM can outperform DMM when $n = 90$ even the latter is the true model. This is because DMM relies on a Laplace approximation to a six dimensional integral when computing the marginal likelihoods to choose the number of clusters. When the sample size is small, DMM tends to choose less than three clusters due to the poor approximation. When $n = 180$, DMM is more likely to choose the right number of clusters even with the inaccurate approximation. Thus the performance of DMM improves significantly with more samples. Our experience suggests that DMM tends to underestimate the number of clusters in most cases. For example, in scenario I

Table 4.3: RMSE of the Jaccard index (small sample size). Cells with the lowest RMSE in each row are highlighted.

$n = 90$									
Expt		Signal		Method					
		Level	R^2	DTM	DMM	K-ms	PAM	Hclust	Spec
I	DT	–	–	**0.43**	0.51	–	–	–	–
		W	0.30	**0.56**	0.64	0.67	0.71	0.65	0.71
		M	0.35	**0.33**	0.65	0.69	0.69	0.64	0.71
		S	0.37	**0.17**	0.65	0.69	0.71	0.65	0.70
II	Dir	–	–	0.35	**0.00**	–	–	–	–
		W	0.35	**0.53**	**0.53**	0.55	0.59	0.58	0.57
		M	0.52	**0.18**	0.30	0.37	0.33	0.37	0.33
		S	0.60	**0.04**	0.09	0.32	0.19	0.38	0.22
III	LN-A	–	–	0.46	**0.06**	–	–	–	–
		W	0.37	**0.49**	0.64	0.53	0.53	0.54	0.54
		M	0.38	**0.23**	0.64	0.50	0.46	0.55	0.47
		S	0.39	**0.10**	0.64	0.48	0.44	0.53	0.46
IV	LN-S	–	–	0.60	**0.54**	–	–	–	–
		W	0.10	**0.35**	0.72	0.77	0.78	0.73	0.74
		M	0.41	**0.21**	0.54	0.59	0.54	0.60	0.53
		S	0.60	**0.17**	0.37	0.36	0.24	0.41	0.27
V	LN-M	–	–	**0.41**	0.61	–	–	–	–
		W	0.04	**0.20**	0.78	0.78	0.79	0.78	0.76
		M	0.23	**0.14**	0.65	0.76	0.70	0.74	0.68
		S	0.53	**0.17**	0.49	0.22	0.20	0.39	0.22

and III, DMM simply puts all samples in a same cluster when $n = 90$.

In our simulation settings, there are two factors that determine the effect of the increase of sample size on the performance of the two model-based clustering methods. On the one hand, since more samples are available per cluster, the models have a better chance to capture the cluster centroids well once they identify the correct number of clusters. On the other hand, more samples makes it harder to get the number of clusters right. These two fighting forces together determine the

Table 4.4: RMSE of the Jaccard index (large sample size). Cells with the lowest RMSE in each row are highlighted.

$n = 180$									
Expt		Signal		Method					
		Level	R^2	DTM	DMM	K-ms	PAM	Hclust	Spec
I	DT	–	–	**0.38**	0.75	–	–	–	–
		W	0.30	**0.45**	0.65	0.67	0.72	0.65	0.71
		M	0.35	**0.36**	0.69	0.69	0.72	0.66	0.71
		S	0.36	**0.19**	0.72	0.69	0.71	0.65	0.70
II	Dir	–	–	0.50	**0.00**	–	–	–	–
		W	0.34	0.51	**0.46**	0.58	0.59	0.59	0.56
		M	0.51	0.16	**0.12**	0.33	0.29	0.37	0.28
		S	0.59	0.08	**0.04**	0.31	0.17	0.39	0.22
III	LN-A	–	–	0.67	**0.21**	–	–	–	–
		W	0.37	0.56	0.55	0.52	**0.48**	0.56	0.51
		M	0.38	**0.26**	0.55	0.49	0.40	0.56	0.44
		S	0.39	**0.13**	0.56	0.46	0.38	0.54	0.41
IV	LN-S	–	–	0.74	**0.66**	–	–	–	–
		W	0.09	**0.62**	0.75	0.77	0.78	0.72	0.73
		M	0.40	**0.46**	0.52	0.58	0.54	0.62	0.47
		S	0.59	**0.19**	0.33	0.34	0.22	0.42	0.20
V	LN-M	–	–	**0.48**	0.69	–	–	–	–
		W	0.03	**0.29**	0.81	0.79	0.79	0.78	0.78
		M	0.22	**0.24**	0.73	0.76	0.71	0.74	0.57
		S	0.52	**0.17**	0.25	0.17	**0.17**	0.24	0.17

overall performance shift of the two model-based methods, yet which force prevails is unclear. For DTMM, when the model is mis-specified (as in scenarios III, IV and V), the model tends to identify too many small clusters, resulting in a worse overall performance. For the distance-based clustering methods, these two factors play no role since we assume that the number of clusters is known. In general, our observations suggest that these methods benefit a little from more samples when the signal is strong. Among the distance-based methods, PAM and Spec have a better

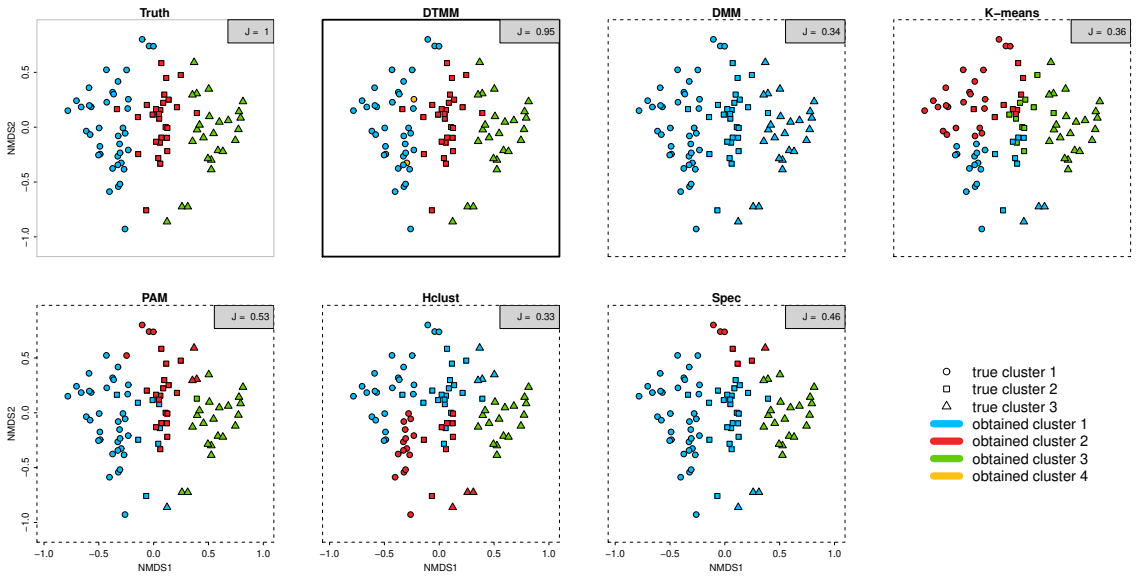

FIGURE 4.10: 2D NMDS plot of samples in a simulation round in scenario IV ($n = 90$, medium noise level). In each sub-plot, the true clustering is indicated by the shape of the points while the clustering obtained is indicated by the color.

overall performance. We thus recommend using these two methods to help choose the initial values of the MCMC chain of DTMM.

We next zoom in to an example to further study the properties of DTMM. In this example, we consider a specific simulation round in scenario IV with the medium noise level ($n = 90$). Figure 4.10 shows the 2D NMDS plot of the samples colored by the clustering obtained by each method. In this example, the clustering is roughly determined only by the first NMDS axis. By utilizing the information provided by the phylogenetic tree, DTMM is capable of picking the relevant dimensions and clustering efficiently. As for a representative clustering, DTMM finds 4 clusters, with one falsely identified cluster containing only two samples. This is consistent with the well-known fact that inference based on Dirichlet process mixture models can identify small clusters that do not reflect the true data-generating process (Miller and Harrison, 2013). One feature that differs DTMM from its competitors is that it not only outputs a

representative clustering, but also a whole MCMC trajectory that allows natural uncertainty quantifications. Figure 4.11 (a) shows the probability of two samples being clustered together by DTMM. Clearly, three stable clusters are identified. Although DTMM falsely puts the first two samples in a separate cluster, the uncertainty is large. There is a high probability that these two samples are actually from the second cluster. The estimated centroids for the three large clusters obtained by DTMM have compositions $(0.44, 0.05, 0.23, 0.22, 0.03, 0.03)$, $(0.24, 0.25, 0.23, 0.22, 0.03, 0.03)$ and $(0.08, 0.41, 0.23, 0.22, 0.03, 0.03)$. In comparison, the centroids' compositions in the true data-generating process are $(0.51, 0.07, 0.19, 0.19, 0.03, 0.03)$, $(0.29, 0.29, 0.19, 0.19, 0.03, 0.03)$ and $(0.07, 0.51, 0.19, 0.19, 0.03, 0.03)$, respectively. A byproduct of DTMM is its ability to perform node selection during the clustering procedure. Figure 4.11 (b) shows the relative abundance of the samples as well as the estimated posterior node selection probabilities. In this example, DTMM is able to uncover the correct subset of internal nodes that are relevant for clustering. We also consider an example from simulation scenario V. Illustrations similar to Figure 4.10 and Figure 4.11 can be found in Figure B.3 and Figure B.4 in the online supplementary materials A.2.

4.3.2 Validation

Validating the results of unsupervised learning is often challenging. In microbiome clustering analyses, the best practice is to check the resulting clusters with scientists to see if they provide any biological insights on a case by case basis. In this section, instead of trying to provide a general solution of how to justify the clusters found by DTMM, we provide an example to show that DTMM can identify biologically meaningful clusters in real microbiome applications.

Specifically, we reanalyze the data in Dethlefsen and Relman (2011), which studies the responses of stable gut microbiota to antibiotic disturbance. In this study, the

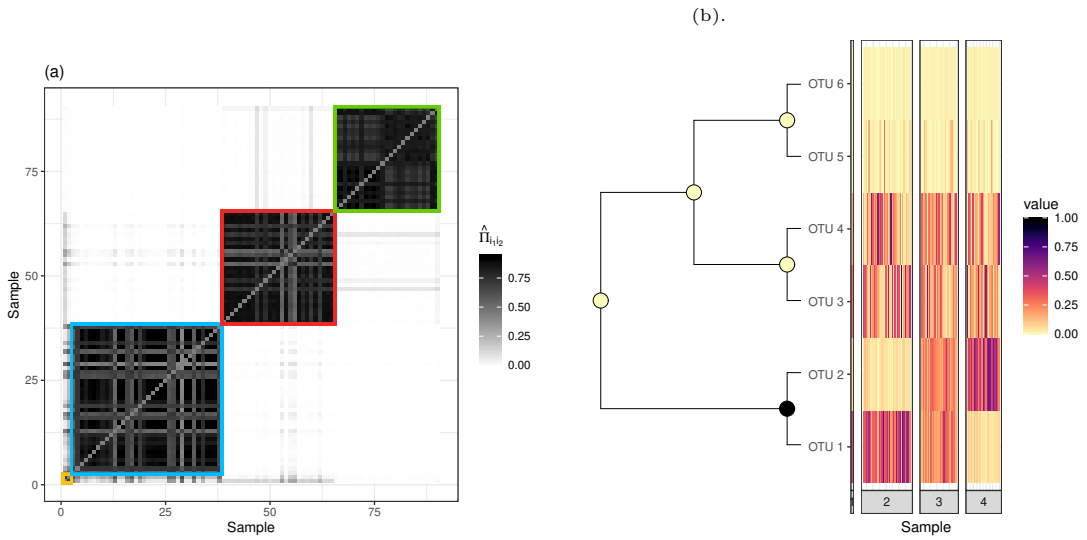

FIGURE 4.11: Illustrations for an example from simulation scenario IV. (a): Probability of two samples being clustered together by DTMM based on 1000 post-burnin MCMC samples. The samples are ordered by their cluster labels from DTMM. The clusters identified by DTMM are highlighted by squares colored as in Figure 4.10. (b): An illustration of the node selection property of DTMM. The nodes are colored by their estimated posterior node selection probabilities. The heatmap plots the relative abundance of the samples grouped by their cluster labels from DTMM.

distal gut microbiome of three patient (patient D, E and F) were monitored over 10 months, including two 5-day antibiotic treatment courses separated by a 5-month interim period. 52 to 56 samples were collected for each patient in the experiment. For example, samples of patient D and F are shown in 4.12 and 4.13. These figures also illustrate the design of the study. In our analyses, the OTU counts are aggregated to the genus level, which gives 59 OTUs (genus) in total.

We analyze the samples from the three patients separately. For each patient, we ignore the time information of when the samples were taken and run DTMM on these samples for 2500 iterations. The first half of the chain was discarded as burn-in. The clustering results for patient D and F are shown in 4.12 and 4.13 (the x-axis labels in these plots are colored by the cluster labels in $\boldsymbol{C}_{LS}$ of the samples they represent). For patient D, DTMM identifies three clusters, which can be interpreted

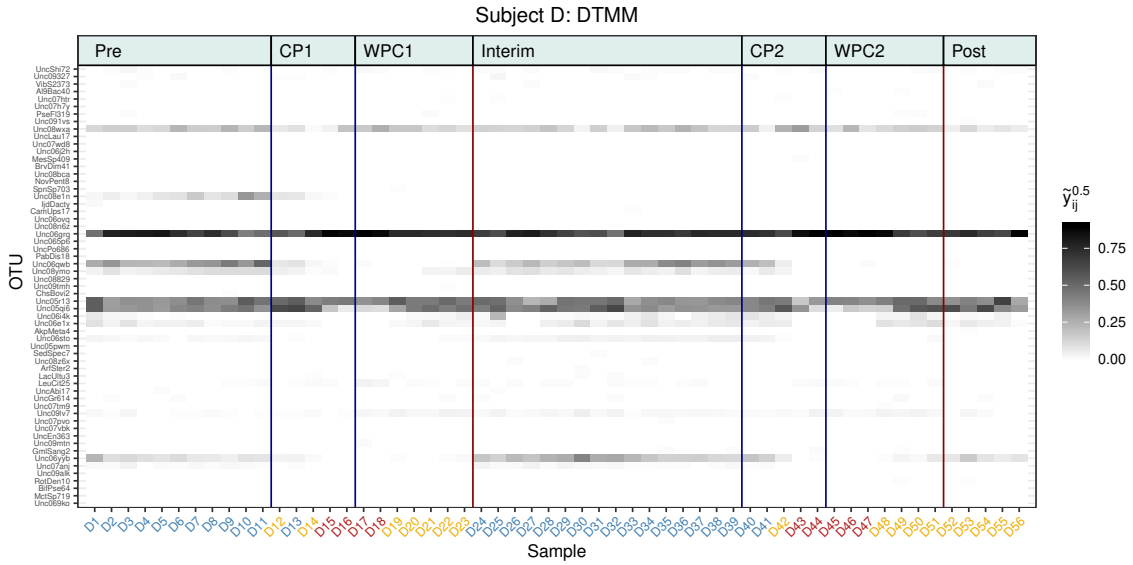

FIGURE 4.12: The heatmap of the microbiome samples of patient D (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DTMM. The blue vertical lines mark the two antibiotic treatment courses. "CP" denotes the antibiotic treatment (ciprofloxacin); "WPC" is the week post treatment; "Pre" and "Post" denote the pre-treatment and post-treatment period, respectively.

as the *stable*, *sterile* and *recover* stages of the microbiota. Based on the clustering results, the gut microbiota of patient D was stable before the treatment. It was able to recover to some stable states from antibiotic treatment within a week after the treatment was finished. However, although the microbiota was able to fully recover to the pre-treatment state after the first antibiotic treatment course, it never made a full recovery to the original state after the second (repeated) antibiotic treatment.

For patient F, DTMM identifies four clusters corresponding to the *stable 1* (blue), *sterile*, *recover* and *stable 2* (purple) stages, respectively. Like patient D, the gut microbiota of patient F was stable before the treatment and was able to recover from the treatments. Unlike patient D, it did not recover to the pre-treatment state even after the first treatment course. Moreover, it took longer for patient F to recover than

patient D. We note that these findings are all consistent to the findings in Dethlefsen and Relman (2011), where the time and design information was used to get these results. Therefore, the clusters found by DTMM are biologically interpretable in this example.

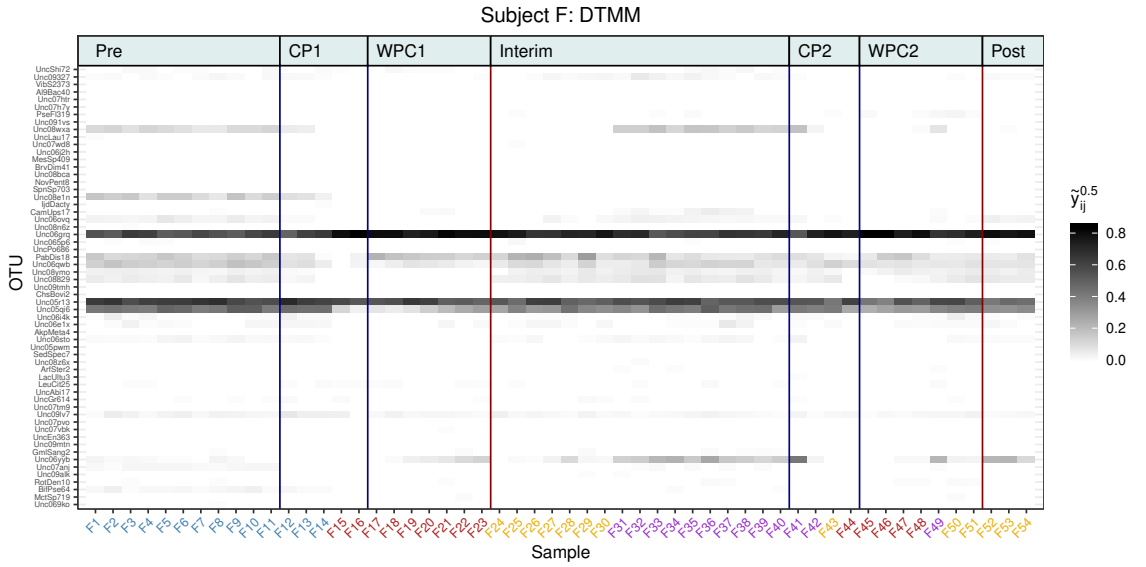

FIGURE 4.13: The heatmap of the microbiome samples of patient F (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DTMM. The legends are defined the same way as in 4.12.

As a comparison, the clustering results for these two patients under DMM are shown in B.5 and B.6 in the online supplementary materials A.2. For both patients, DMM returns two clusters with high probability, with the two clusters roughly representing the *stable* and *unstable* stages of the microbiota, respectively. In this example, DTMM is able to discover more interesting latent structures among samples than DMM. It is worth noting that in each analysis, microbiome samples were collected from the same patient. Thus the level of cross-sample variations in this study is much smaller than microbiome studies with multiple subjects. In those cases, we expect DTMM to benefit more from its improved flexibility over DMM and discover

even more interpretable structures than the latter.

4.4 Application to the American Gut project data

In this section, we apply DTMM to the July 2016 version of the fecal data from the AG project to construct enterotypes for two groups of samples: firstly, we consider participants who have been diagnosed with inflammatory bowel disease (IBD); secondly, we consider participants who have been diagnosed with diabetes. The diagnoses are made by a medical professional (a doctor or a physician assistant). We focus on the top 75 OTUs based on total counts to reduce noises in the dataset and control for the sequencing errors. The top 75 OTUs on average retain 2/3 of the total counts in a sample. We filter the samples by only considering participants with at least 500 counts on the top 75 OTUs. This filtering ends up with 189 samples diagnosed with IBD and 106 samples diagnosed with diabetes that we use in our analysis.

In the following sections, we fit DTMM to each or the two datasets with the priors and hyperparameters set to the recommended choices in Section 4.2.3. In each analysis, we run the Gibbs sampler in Section 4.2.4 for 5000 iterations and discard the first half of the chain as burn-in. The cluster labels are initiated by running the PAM algorithm with $K = 5$.

Key findings from our analyses are summarized as follows:

- The enterotypes (clusters) are determined by a large number of OTUs jointly in a sophisticated manner instead of by a few OTUs.

- OTUs from genera *Bacteroides, Prevotella* and *Ruminococcus* are typically important in identifying enterotypes, which is consistent to the findings in previous works (Arumugam et al., 2011).

- The number of enterotypes and the OTUs that characterize each enterotype can differ across datasets.
- DMM tends to find larger clusters that are unions of clusters found by DTMM.

4.4.1 IBD

We first consider samples from participants that are diagnosed with IBD. Figure 4.14 shows the traceplots of some one dimensional parameters or summaries of the posterior samples. The Markov chain stabilizes and mixes reasonably well after about 750 iterations. Figure 4.14 (a) and Figure 4.14 (b) show the traceplots of the Dirichlet process precision parameter β and the prior global coupling probability λ. The posterior means of these parameters are 0.87 and 0.53, respectively. Traceplot of the sampled number of internal nodes with $\gamma(A) = 1$ is shown in Figure 4.14 (c). On average, 39 out of 74 nodes are marked as relevant to the clustering process, indicating that the clustering process is determined by various OTUs jointly in a complicated way instead of being dominated by a few OTUs that are dominant in terms of counts. Figure 4.14 (d) shows the cumulative proportion of samples in the largest one, two, three, four and five clusters for each iteration. DTMM tends to assign samples into 5 clusters.

We find $\boldsymbol{C}_{LS}$ as defined in (4.29), which corresponds to $\boldsymbol{c}^{(t_0)}$ with $t_0 = 2717$. $\boldsymbol{C}_{LS}$ assigns the samples into five clusters, with size 6, 41, 73, 42 and 27, respectively. The estimated centroids of the five clusters are shown in Figure 4.15. Figure 4.15 also shows the estimated posterior means of the coupling indicator $s(A)$ at $A \in \mathcal{I}$. Most internal nodes that are irrelevant to clustering are close to the leaves of the tree. Nodes that are more "global" (have more descendant OTUs) generally contribute to the clustering. This indicates that the clustering process is determined by most OTUs jointly in a complicated manner. Figure 4.16 (left) shows the estimated pairwise clustering probability matrix $\hat{\Pi}$ with the rows and columns ordered by the labels in

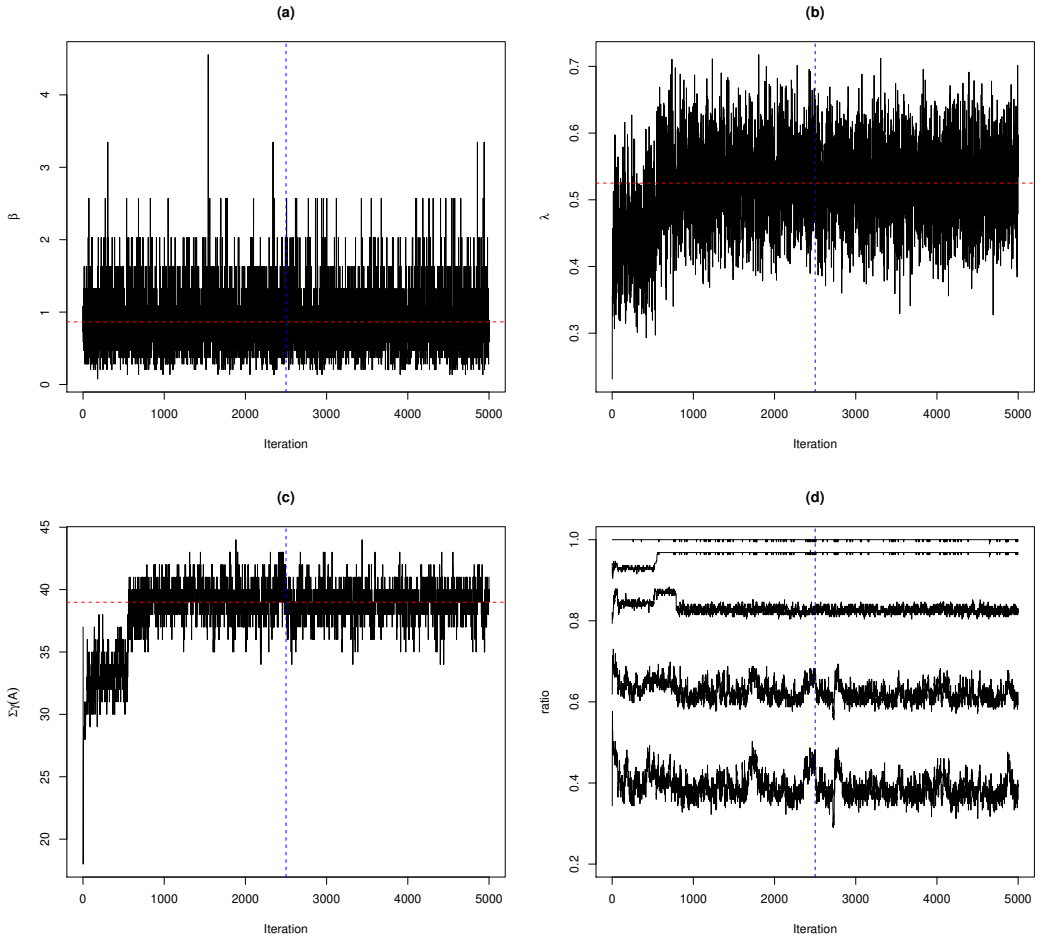

FIGURE 4.14: Traceplots of some summary statistics.

$\boldsymbol{C}_{LS}$. There are noticeable uncertainties in the clustering, especially between cluster 2, 3 and 4. This can also be seen from Figure 4.16 (right), where we plot the heatmap of the samples (after the square-root transform) grouped by their labels in $\boldsymbol{C}_{LS}$. Clearly, clusters 2, 3 and 4 are similar. Figure 4.16 (right) also shows that the within-cluster variations among samples are large.

To see which OTUs are more important in determining $\boldsymbol{C}_{LS}$, we consider the

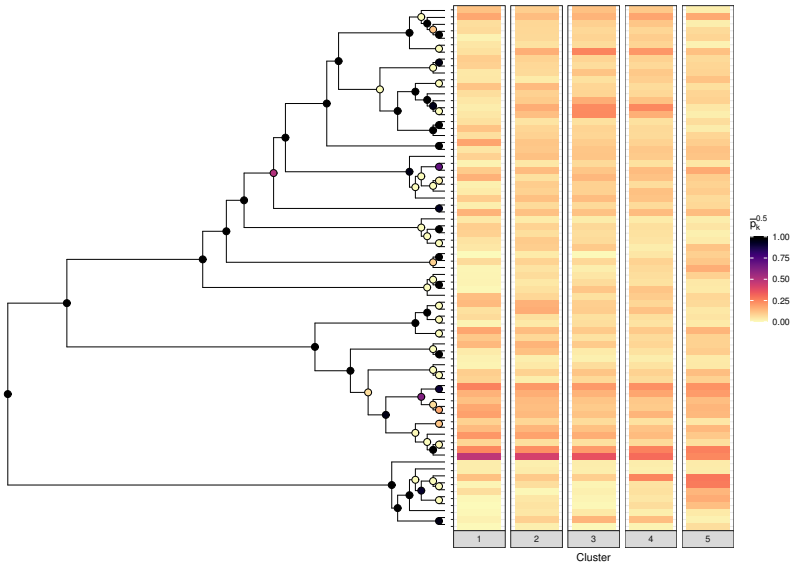

FIGURE 4.15: Left: Estimated posterior means of the coupling indicators at each node of the phylogenetic tree. Right: The estimated centroids of the five clusters in $\boldsymbol{C}_{LS}$ (after the square-root transform).

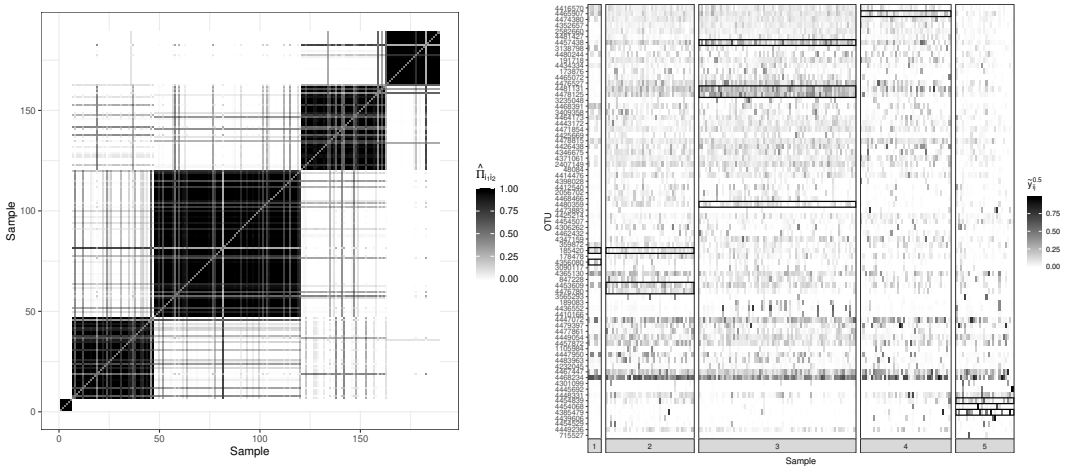

FIGURE 4.16: Left: Estimated pairwise co-clustering probabilities. Right: Heatmap of the samples (after the square-root transform) grouped by their labels in $\boldsymbol{C}_{LS}$. The black boxes illustrate the characteristic OTUs of each cluster.

following heuristic measure of OTU importance: for $1 \leqslant j \leqslant M$, let

$$\vartheta_j = \frac{SSB_j}{SSW_j} = \frac{\sum\limits_{c \in \boldsymbol{C}_{LS}} n_c (\bar{y}_{cj} - \bar{y}_j)^2}{\sum\limits_{c \in \boldsymbol{C}_{LS}} \sum\limits_{c_i = c} (y_{ij} - \bar{y}_{cj})^2}, \tag{4.36}$$

Table 4.5: Estimated cluster-specific compositions of the top 10 OTUs in determining $\boldsymbol{C}_{LS}$ in terms of ϑ_j. Values of the OTU compositions are shown in the percentage scale.

OTU	Family	Genus	ϑ_j	C1	C2	C3	C4	C5
185420	Bacteroidaceae	Bacteroides	0.43	2.01	2.62	0.98	0.66	0.54
4478125	Ruminococcaceae	Faecalibacterium	0.43	0.22	1.71	5.72	2.71	0.10
4356080	Barnesiellaceae	–	0.33	0.53	0.41	0.37	0.42	0.28
4476780	Rikenellaceae	–	0.32	0.14	1.70	0.15	0.33	1.04
4453609	Rikenellaceae	–	0.26	2.08	2.43	1.15	0.71	0.85
4480359	Ruminococcaceae	–	0.22	0.22	1.02	1.22	0.11	1.48
4465907	Lachnospiraceae	Blautia	0.21	3.32	1.82	2.40	3.47	3.04
4481131	Ruminococcaceae	Faecalibacterium	0.18	0.11	2.92	5.62	6.11	0.22
4457438	Lachnospiraceae	–	0.19	0.36	2.72	6.43	4.67	1.68
4385479	Enterobacteriaceae	Proteus	0.17	0.02	0.21	0.04	0.24	2.91

where $\bar{y}_j$ is the overall mean of y_{ij}, $\bar{y}_{cj}$ the mean of y_{ij} for samples with $c_i = c$. Table 4.5 shows the top 10 OTUs in determining $\boldsymbol{C}_{LS}$ in terms of ϑ_j as well as their compositions in each cluster centroid. Overall, $\boldsymbol{C}_{LS}$ is jointly determined by multiple OTUs in a complicated way. Note that OTUs that are important for clustering are not necessarily those with abundant counts. For example, OTU-4468234 and OTU-4447072 (both are *Bacteroides*) are the two OTUs with the most counts in the dataset. However, these two OTUs are prevalent in most samples and thus have limited roles in the clustering.

We next compare the five resulting clusters in more details. Figure 4.17 shows the boxplot of the Shannon diversity of samples in the five clusters, respectively. Samples from cluster 2 and 3 tend to have more evenly distributed counts across OTUs compared to those from cluster 1 and 5. Similar to (4.36), we can define a heuristic measure of OTU importance in characterizing each of the five clusters. Specifically, for $c = 1, \ldots, 5$, let

$$\vartheta_j^c = \frac{SSB_j^c}{SSW_j^c} = \frac{n_c(\bar{y}_{cj} - \bar{y}_j)^2 + n_{-c}(\bar{y}_{-cj} - \bar{y}_j)^2}{\sum\limits_{c_i=c}(y_{ij} - \bar{y}_{cj})^2 + \sum\limits_{c_i \neq c}(y_{ij} - \bar{y}_{-cj})^2}, \tag{4.37}$$

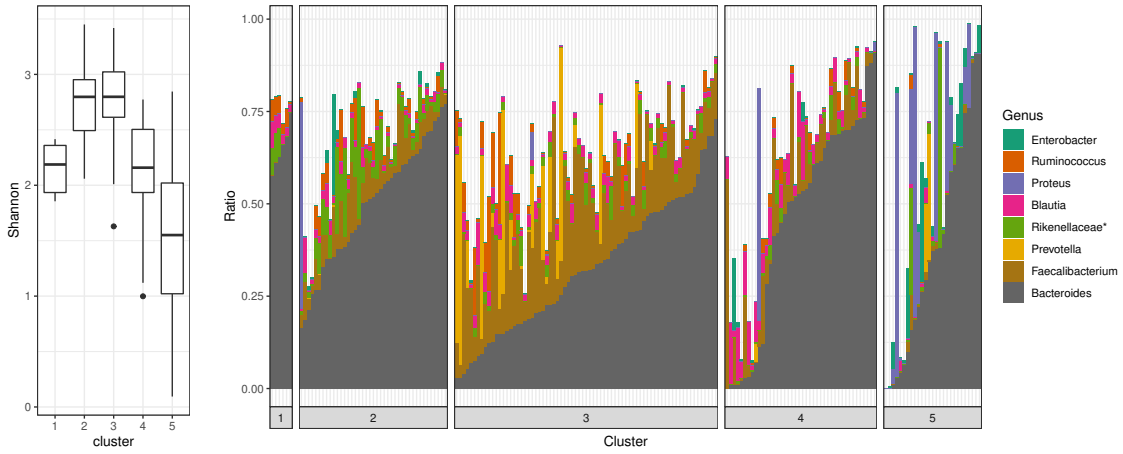

FIGURE 4.17: Left: Boxplot of the Shannon diversity of samples in each cluster. Right: Relative abundance of 8 genera for each sample. A genus is chosen if its descendant OTUs have large ϑ_j^c for some c. For the 3 OTUs with unavailable genera information, their family is shown instead (indicated by *Rikenellaceae**).

where n_{-c} is the number of samples that are not in cluster c, $\bar{y}_{-cj}$ the mean of y_{ij} for samples with $c_i \neq c$.(4.37) is equivalent to merging the four clusters other than cluster c in (4.36). The boxes in Figure 4.16 (right) indicate the top OTUs in terms of ϑ_j^c for each c (only OTUs with $\vartheta_j^c > 0.1$ are shown).

Based on these results, we can characterize each cluster by a few OTUs with the top ϑ_j^c. For example, samples from cluster 2 tend to have more counts from the *Rikenellaceae* family (represented by OTU-4453609 and OTU-4476780). Cluster 3 is characterized by having more abundance in the *Faecalibacterium* (represented by OTU-4478125 and OTU-4481131) and the *Lachnospiraceae* family (represented by OTU-4481127 and OTU-4457438). Arumugam et al. (2011) proposed three enterotypes in human gut microbial communities that are characterized by the variation in the levels of one of the three genera: *Bacteroides, Prevotella* and *Ruminococcus*. Our analysis suggests that the enterotypes are determined by a more sophisticated mechanism involving more genera. Although OTUs from the *Bacteroides, Prevotella*

and *Ruminococcus* genera are not always those with the largest ϑ_j^c, they are playing important roles in identifying the five clusters. For example, OTUs from the *Prevotella* genus have large ϑ_j^3 and are thus crucial in determining cluster 3 while OTUs from the *Ruminococcus* genus have large ϑ_j^1 and ϑ_j^2 and are thus important in identifying cluster 1 and 2. This can also be seen from Figure 4.17 (right), where relative abundance of 8 genera picked by ϑ_j^c are shown for each sample. The samples are grouped by their cluster labels in $\boldsymbol{C}_{LS}$.

4.4.2 Diabetes

Similar to Section 4.4.1 we apply DTMM to samples from participants that are diagnosed with diabetes. Counterparts of Figure 4.14 and Figure 4.16 are shown in Figure B.7 and Figure B.8 in the online supplementary materials C. In this example, DTMM finds three clusters with $\boldsymbol{C}_{LS}$. Figure 4.18 shows the estimated centroids of the three clusters as well as the estimated posterior means of the coupling indicators at each node of the phylogenetic tree. Figure 4.17 (right) shows for each sample the relative abundance of 6 genera selected based on the importance of their descendant OTUs in identifying the three clusters. Compared with the IBD example, the enterotypes in this case can be identified easier. For example, samples in cluster 3 tend to have significantly lower abundance in *Faecalibacterium* and *Bacteroides*, which are the dominating genera in most samples. Compared with cluster 2, cluster 1 is identified with relatively more counts from OTU-173876 and the *Prevotella* family.

On average, 21 out 75 internal nodes of $\mathcal{T}$ are estimated as relevant to the clustering process. Compared with the IBD example in Section 4.4.1, fewer nodes are involved. The reason for this is twofold. Firstly, as we discussed previously, clusters in the diabetes example are determined by fewer OTUs (genera), which is reflected in the number of internal nodes selected. Secondly, as shown in Figure B.8 (right), a few OTUs play crucial roles in determining multiple clusters. As a comparison,

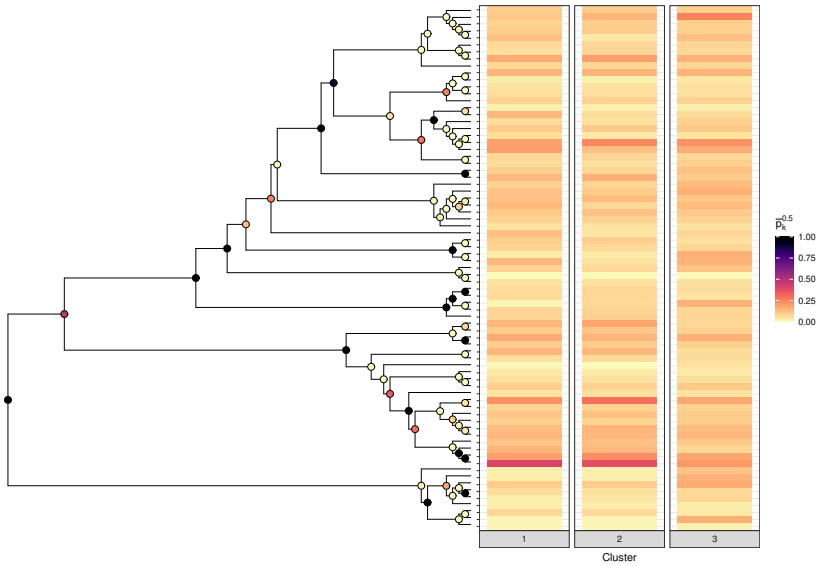

FIGURE 4.18: Left: Estimated posterior means of the coupling indicators at each node of the phylogenetic tree. Right: The estimated centroids of the three clusters in $\boldsymbol{C}_{LS}$ (after the square-root transform).

as shown in Figure 4.16 (right), each cluster in the IBD example is determined by a unique set of OTUs. Since DTMM marks an internal node as relevant if it is relevant in determining *any* cluster, more nodes are selected in the IBD example.

4.4.3 DTMM vs. DMM

We also apply DMM to the two examples in this section and compare it with DTMM. DMM reports two clusters in both examples (see Figure B.10 in the online supplementary materials C). For the IBD dataset, Figure 4.20 (left) shows the two-dimensional NMDS plot of the data, colored by the cluster labels reported by DMM. In comparison, Figure 4.20 (right) shows the same NMDS plot colored by the cluster labels in $\boldsymbol{C}_{LS}$ reported by DTMM.

The five clusters reported by DTMM can be seen as refinements of the two clusters reported by DMM. Roughly, cluster B identified by DMM is the union of cluster 2

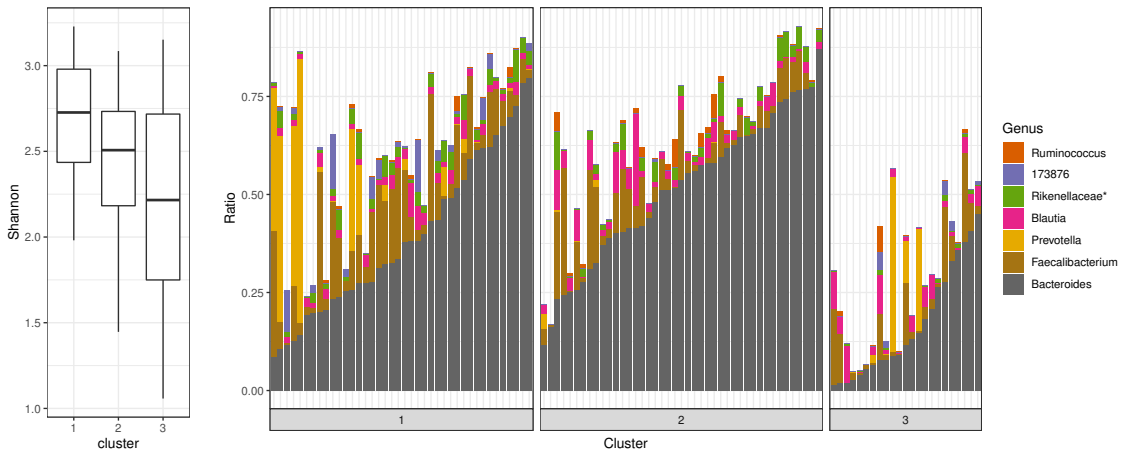

FIGURE 4.19: Left: Boxplot of the Shannon diversity of samples in each cluster. Right: Relative abundance of 6 genera for each sample. A genus is chosen if its descendant OTUs have large ϑ_j^c for some c. For the 3 OTUs with unavailable genera information, their family is shown instead (indicated by *Rikenellaceae**). OTU-173876 is important in identifying cluster 1 and is illustrated directly since no genera or family information is available for this OTU.

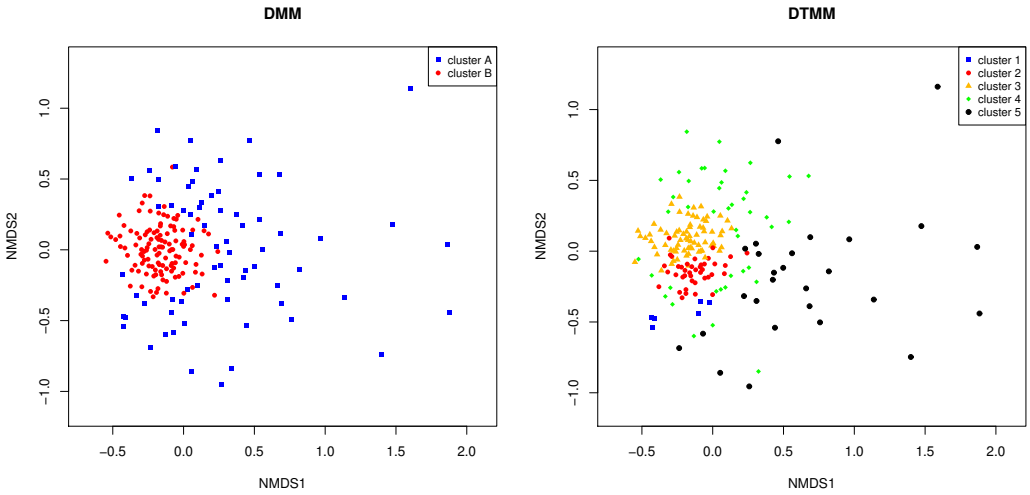

FIGURE 4.20: Two-dimensional NMDS plots for the AG IBD dataset. Left: Points are colored and shaped by the resulting cluster given by DMM. Right: Points are colored and shaped based on $\boldsymbol{C}_{LS}$.

and 3 from DTMM while cluster A found by DMM is the union of cluster 1, 4 and 5 from DTMM. As shown in Figure 4.16 (right) and Figure 4.17 (right), those subclusters from DTMM are not differentiated by OTUs with dominant counts. Thus it is very unlikely for DMM to make further splits. Moreover, based on Figure 4.16 (right), samples within each cluster from DTMM tend to show different levels of heterogeneities across OTUs, making the underlying Dirichlet-multinomial model of DMM unrealistic. For example, counts of OTUs from the *Prevotella* genus tend to show large within cluster variations among samples in cluster 3. To capture this level of variation, DMM has to push the cluster-specific dispersion parameter very large, essentially loose its ability to effectively find those subclusters.

4.5 Concluding remarks

We have introduced DTMM as a model-based framework for clustering the amplicon sequencing data in microbiome studies. By directly incorporating the phylogenetic tree, DTMM generalizes the popular DMM in three directions: first, it offers a more flexible covariance structure among different OTUs; second, it provides a way for selecting a subset of internal nodes in the phylogenetic tree that is relevant for clustering; moreover, it replaces the finite mixtures in DMM with Dirichlet process mixtures to allow automatic selection of the number of clusters.

Although the covariance structure offered by DT is richer than that of the Dirichlet distribution, it is still limited compared to the logistic-normal family (LN). In a case with K OTUs, DT models the covariance among OTU counts with $(K-1)$ dispersion parameters in the series of beta distributions while LN uses $K(K-1)/2$ parameters in modeling the covariance matrix. It is interesting to further generalize the covariance structure provided by DTMM without making the inference too complicated. When selecting a subset of internal nodes in the phylogenetic tree that are relevant to clustering, DTMM selects a node if it is relevant in identifying *any*

cluster. Intuitively, DTMM first selects a subspace in the node space and performs clustering in that space. An alternative direction worth exploring is to allow the nodes selected to be cluster-dependent such that each cluster can deviate from the "mean" cluster at different internal nodes.

Based on our analysis of the AG data, enterotypes are defined by a large number of OTUs jointly in a complex manner. Other than the dominant OTUs, some OTUs that are not very abundant in counts are also crucial in determining these clusters.

5

Bayesian recursive partitioning for causal inference

5.1 Introduction

Making unconfounded comparisons between groups of samples and determining the effect of a treatment is a central task in scientific research. A gold standard for such comparisons is through carefully designed randomized experiments (Rubin, 2008a), in which experimental units are randomly assigned into treatment groups under a known mechanism. For example, in completely randomized experiments, directly comparing units in different treatment groups has valid causal interpretations by virtue of randomization. In observational studies, such direct comparison introduces systematic bias in estimating the treatment effects due to the unknown treatment assignment mechanism and the potential "unbalance" of relevant pre-treatment covariates between the treatment groups, that is, their distributions in these groups are substantially different.

An example of randomized design is the randomized block design (RBD), in which experimental units are arranged into blocks according to certain blocking factors, and units within each block are assigned to different treatment groups completely at ran-

dom. Within each block, the effect of the nuisance blocking factors is removed and the effect of the treatment of interests can be estimated through directly comparing units in different treatment groups. Sometimes, a properly designed observational study could be viewed as an approximation to an underlying hypothetical RBD (Rubin, 2008b). A popular way for valid causal comparison is to uncover this underlying RBD and compare units in different treatment groups conditioning on the uncovered blocks. For example, Figure 5.1 shows the two-dimensional covariates of units in a fictional observational study. Within each of the four blocks, this observational study approximates a completely randomized design.

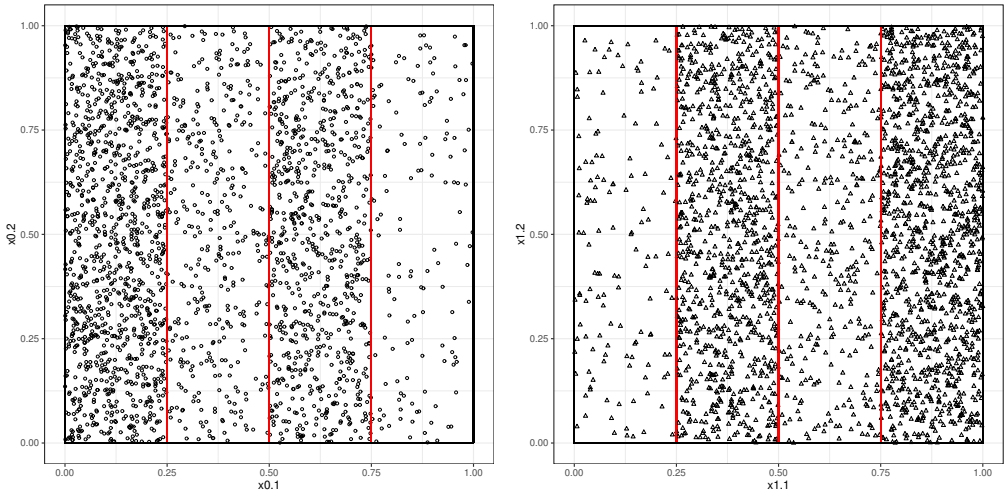

FIGURE 5.1: A fictional observational study that approximates an RBD with four blocks. Left: covariates of units in the control group. Right: covariates of units in the treatment group.

Following the seminal work of Rosenbaum and Rubin (1983), adjusting for the propensity score is sufficient to reduce the systematic bias in estimating the causal effect induced by confounding under the strong ignorability assumption. Therefore, arranging units with similar propensity scores provides a convenient way of uncovering an underlying RBD design in an observational study, referred to as the propensity score stratification (Rosenbaum and Rubin, 1984). This classic solution

only requires estimating the one-dimensional propensity score for each unit, which seems to be much easier than dealing with the multidimensional covariate and trying to find balancing blocks in the covariate space directly. Over the decades, various off-the-shelf tools have been used to estimate the propensity score, ranging from the simple logistic regression (Rosenbaum and Rubin, 1983) to the flexible nonparametric generalized boosting models (McCaffrey et al., 2004).

The propensity score stratification falls into a family of methods, generally referred to as "matching", that aim at balancing the distribution of the pretreatment covariates in different treatment groups (Stuart, 2010). Other methods in this family include nearest-neighbor matching (Rosenbaum and Rubin, 1985; Rosenbaum, 1989; Hansen, 2004), weighting (Hirano et al., 2003; Crump et al., 2009; Li et al., 2017) and variants of them. It is worth noting that for these propensity score based methods, the propensity score itself is not the variable of interest *per se*, but a tool through covariate balancing. Thus a propensity score estimation can only be evaluated based on its ability in balancing the covariates. For example, in the propensity score stratification, if in some blocks the covariates are not well-balanced, the propensity score must be estimated again by another model. In practice, this procedure might need to be carried out multiple times before a reasonable blocking is achieved, which can be inefficient and has no clear rule as to when this loop should stop.

Other than this "propensity score tautology" (Imai et al., 2008), these methods have other limitations. For example, they usually entail *ad hoc* tuning parameters, such as the number of matched blocks in the stratification procedure (Rosenbaum and Rubin, 1984) and the number of control unites matched to each treatment units in nearest-neighbor matching (Smith, 1997). In addition, units with similar propensity scores can be far away in the original covariate space. Yet they are likely to be matched together by the propensity score based methods. This is counterintuitive and undesirable when the treatment effect is not homogeneous.

In this chapter, our first goal is to introduce a method that partitions the covariate space into balancing blocks directly, without resorting to the propensity score. To this end, we take a retrospective perspective on the treatment status and model the covariates of units in different treatment groups jointly. At the first glance, this seems to be "impossible" compared to estimating the propensity score. However, our goal here is not to accurately learn the multivariate covariate distribution, but the induced balancing blocks, which correspond to only a single aspect of the joint covariate distribution. As we shall see, our strategy shares the motivation of learning correlations of variables through copulas. The success of our approach hinges critically on the ability of the modeling strategy to directly model the *relation* of the covariate distributions in different treatment groups instead of modeling them separately and infer their relation from the joint model. By directly targeting at the covariate space, the proposed method finds balancing blocks that are interpretable subpopulations. Moreover, the balancing is measured in a probabilistic manner and the number of blocks needed is determined automatically.

An important feature of the aforementioned strategies for causal comparison is that the *design* stage and the *analysis* stage of the study are separated. For example, in propensity score stratification, balancing blocks can be constructed to reflect the design of the experiment even before seeing the outcome variable. On one hand, this two-staged strategy only utilizes the covariate information and the treatment assignment information to uncover the underlying design, and is thus robust to the specification of an outcome model in the analysis stage. On the other hand, the uncovered design says nothing about the potential heterogeneity of the treatment effect. Following up the fictional example in Figure 5.1, suppose that the outcomes of the experimental units are observed and shown as in Figure 5.2. In this example, the treatment effect is heterogeneous across units and is determined only by the second covariate. Although the four blocks in Figure 5.1 allows unfounded causal

comparison within each block, they do not reflect the underlying treatment effect heterogeneity. In comparison, within each of the 16 blocks shown in Figure 5.2, the covariates are balanced and the treatment effect is homogeneous.

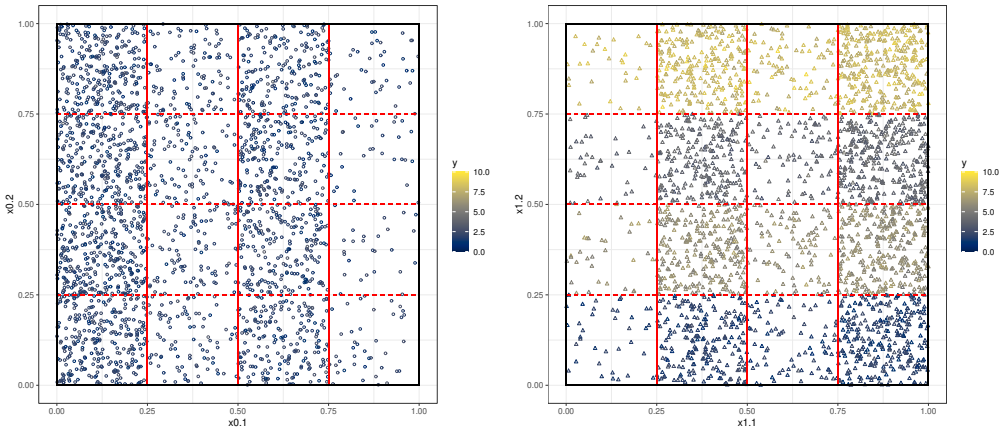

FIGURE 5.2: A fictional observational study that approximates an RBD with four blocks. Left: covariates of units in the control group. Right: covariates of units in the treatment group.

Our second goal in this chapter is to propose a probabilistic way to infer such blocks through incorporating an outcome model. Specifically, we aim at partitioning the samples into blocks within each of which the covariate distributions are balanced *and* the treatment effect is homogeneous. To achieve this, we model the covariates and the outcome jointly conditioning on the treatment status by using the flexible covariate model discussed above and incorporating an outcome model given the covariates. The outcome model allows the partitioning procedure to be efficient and makes it possible for the blocks to capture the heterogeneities in the treatment effects. The covariate model checks covariate balance during the procedure thus protects the outcome model from severe extrapolation.

In the past decade, an emerging literature focuses on addressing the treatment effect heterogeneity (Imai et al., 2013; Tian et al., 2014; Hill, 2011; Athey and Imbens, 2016; Wager and Athey, 2018; Hahn et al., 2020). We note that although the

design stage is not explicitly uncovered in this stream of methods, it is still critically important. For example, even a flexible outcome model as BART can yield badly biased estimates of treatment effects when directly fitted to the outcome data under strong confounding (Hahn et al., 2020). To remove such biases, the design stage has to be taken into account by incorporating the estimated propensity score in the BART model (Hahn et al., 2020). Similarly, in constructing a causal random forest that gives consistent causal effect estimation, an "honest" condition has to be satisfied by each single tree that separates the design and the analysis stage explicitly (Wager and Athey, 2018). In out proposed method, the role of both stages are equally emphasized.

The rest of the paper is organized as follows. In Section 5.2, we introduce a general retrospective framework to achieve joint modeling of the design stage (covariates) and the analysis stage (outcome) in an observational study. In Section 5.3, we study the two-step approach in details, evaluating the performance of the proposed method numerically and from a theoretical perspective.

5.2 The retrospective framework

5.2.1 Set-up

We start with the common setting with two treatment status. Consider a finite sample of size n drawn randomly from some large population. The i-th observation consists of $(\boldsymbol{X}_i, Z_i, Y_i)$ where $\boldsymbol{X}_i \in \Omega_{\boldsymbol{X}}$ is a covariate vector of p dimension, $Z_i \in \{0, 1\}$ the treatment indicator, and Y_i is an outcome. We use $\mathcal{X}_n$ to denote the covariate of all the units and define $\mathcal{Z}_n$ and $\mathcal{Y}_n$ in a similar manner. Moreover, let $\mathcal{D}_n = \{\mathcal{X}_n, \mathcal{Z}_n, \mathcal{Y}_n\}$. Under the standard stable unit treatment value assumption (SUTVA) (Imbens and Rubin, 2015), we take the potential outcome framework to make causal comparisons (Rubin, 1974, 1978) and let $\boldsymbol{Y}_i = \{Y_i(0), Y_i(1)\}$ be the two potential outcomes under control and treatment for unit i. Following Rosenbaum and Rubin

(1983), we also assume that the treatment assignment is strongly ignorable (SI), which requires (i) the unconfoundedness assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i \mid \boldsymbol{X}_i$ and (ii) the probabilistic assignment assumption: $0 < \delta < \Pr(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) < 1 - \delta < 1$ for all $\boldsymbol{x}$ and some $0 < \delta < 1$.

Let $\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}]$ be the conditional average treatment effect (CATE) given $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$. Our goal is to estimate $\tau(\boldsymbol{x})$, which also serves as the basis for estimating population level treatment effects such as the average treatment effect (ATE) or the average treatment effect on the treated (ATT):

$$\tau^{\text{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}_{\boldsymbol{X}}[\tau(\boldsymbol{X})]$$
$$\tau^{\text{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1] = \mathbb{E}_{\boldsymbol{X}|Z=1}[\tau(\boldsymbol{X})].$$

In this work, our goal is to infer a partition of the covariate space such that within each block of the partition, direct comparisons of samples from different treatment groups have causal interpretation. Formally, we are interested in inferring a balancing partition:

Definition 1 (Balancing partition)**.** Let $\mathcal{P}_b = \{B_k : k \in I_b\}$ be a partition of the covariate space $\Omega_{\boldsymbol{X}}$. We shall refer to $\mathcal{P}_b$ as a balancing partition if $Z \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{X} \in B_k$.

Within each block B_k, the covariate distributions under the two treatment regimes are balanced and the treatment assignment imitates a randomized design in controlled experiments. Given $\mathcal{P}_b$, unbiased estimators of τ^{ATE} and τ^{ATT} can be obtained by properly weighting the differences in means of the outcome between the two treatment groups in each block. Let $e(\boldsymbol{x}) = \Pr(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x})$ be the propensity score. Following Rosenbaum and Rubin (1983), a propensity score partition $\mathcal{P}_e$ is a balancing partition:

Definition 2 (Propensity score partition)**.** Let $\mathcal{P}_e = \{E_k : k \in I_e\}$ be a partition of the covariate space $\Omega_{\boldsymbol{X}}$. We shall refer to $\mathcal{P}_e$ as a propensity score partition if (i)

$e(\boldsymbol{x}_1) = e(\boldsymbol{x}_2)$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in E_k$ and (ii) $e(\boldsymbol{x}_1) \neq e(\boldsymbol{x}_2)$ for any $\boldsymbol{x}_1 \in E_{k_1}$, $\boldsymbol{x}_2 \in E_{k_2}$ and $k_1 \neq k_2$.

Specifically, we require the index sets I_b and I_e to be finite. In practice, such $\mathcal{P}_b$ and $\mathcal{P}_e$ may not exist. In this case, our goal is to find an approximation to these partitions. Inferring (an approximation to) $\mathcal{P}_e$ corresponds to the classic propensity score stratification strategy (Rosenbaum and Rubin, 1984). Moreover, since the propensity score is the coarsest balancing score, $\mathcal{P}_b$ is finer than $\mathcal{P}_e$.

Let F denote the marginal distribution of the covariates, F_0 the conditional distribution of the covariates among the controls, and F_1 that among the treated. We also assume that F, F_0 and F_1 have densities $f(\boldsymbol{x}), f_0(\boldsymbol{x})$ and $f_1(\boldsymbol{x})$, respectively. We note that a fully probabilistic model for $\boldsymbol{X}_i$ given Z_i will allow proper inference on the propensity score. Specifically, we define the covariate density ratio (CDR) between two treatment groups as

$$r(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \propto \frac{f(\boldsymbol{x})e(\boldsymbol{x})}{f(\boldsymbol{x})(1-e(\boldsymbol{x}))} = \frac{e(\boldsymbol{x})}{1-e(\boldsymbol{x})}. \tag{5.1}$$

Accordingly, effective estimation of the CDR $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ will allow efficient inference on the propensity score. Specifically, points in the covariate space that have the same CDR also have the same propensity score. This allows a strategy to learn the propensity score partition $\mathcal{P}_e$ (and thus a balancing partition) through learning a CDR partition:

Definition 3 (Covariate density ratio partition)**.** Let $\mathcal{P} = \{A_k : k \in I\}$ be a partition of the covariate space $\Omega_{\boldsymbol{X}}$. We shall refer to $\mathcal{P}$ as a covariate density ratio partition if $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) = c_A$ for any $A \in \mathcal{P}$, where $\delta < c_A < M$ for $\delta, M > 0$.

Due to SI, the joint likelihood of the complete data factorizes as follows

$$\prod_{i=1}^{n} f(Z_i, \boldsymbol{X}_i, Y_i) = \prod_{i=1}^{n} f(\boldsymbol{X}_i) f(Z_i \mid \boldsymbol{X}_i) f(Y_i \mid \boldsymbol{X}_i). \tag{5.2}$$

Most classic methods thus proceed by conditioning on the covariates and directly specifying either the propensity score model $f(Z_i \mid \boldsymbol{X}_i)$ or the outcome model $f(Y_i \mid \boldsymbol{X}_i)$ or both. In this work, we take a retrospective perspective on the treatment status and base the inference on a different likelihood factorization:

$$\prod_{i=1}^{n} f(Z_i, \boldsymbol{X}_i, Y_i) = \prod_{i=1}^{n} f(Z_i) f(\boldsymbol{X}_i \mid Z_i) f(Y_i \mid \boldsymbol{X}_i). \tag{5.3}$$

Specifically, our inference strategy is based on the conditional likelihood given the treatment status, not the covariates.

5.2.2 A Bayesian nonparametric model for covariate density ratios

To learn the CDR and thus the partition $\mathcal{P}$, we take a nonparametric modeling approach, without assuming any parametric forms of the two densities f_0 and f_1. Many nonparametric models are available, but since our objective is not in learning the two densities *per se* but to get an equivalent partition on $\Omega_{\boldsymbol{X}}$ induced by their ratio, we adopt a model under which the CDR is a piecewise constant function. Note that the underlying densities do not have to be piecewise constant but their ratio. This is important as inferring details of the covariate distributions is often harder than inferring their ratio. The piecewise constant assumption on the CDR is essentially nonparametric as well, because any bounded CDR can be approximated arbitrarily well by piecewise constants. (The boundedness of the CDR is in essence the probabilistic assignment assumption.) Such a model will be particularly effective in achieving local smoothing on the CDR and hence the propensity score.

In summary, we construct a joint Bayesian nonparametric model on the pair of covariate distributions (F_0, F_1) (or equivalently, the covariate densities (f_0, f_1)) such that under this model,

(i) f_0 and f_1 can both be any bounded densities;

(ii) $r = f_1/f_0$ is approximated by a piecewise constant function $r_{\mathcal{P}}(\boldsymbol{x}) = \sum_{A\in\mathcal{P}} c_A \cdot \mathbb{1}_{\boldsymbol{x}\in A}$ with finite pieces, and can be effectively inferred upon while providing full uncertainty quantification. Note that $\mathcal{P}$ is the induced CDR partition.

In practice, we shall require each block in the CDR partition to be contiguous (i.e., connected through a path within the block). While the true propensity score partition may be disconnected, allowing such distconnected blocks makes the model space on the CDR or propensity score too flexible and can introduce substantial variance into the inference.

Our model takes the following hierarchical form:

$$\begin{aligned} \mathcal{P} &\sim \pi(\mathcal{P}) \\ (F_0, F_1) \mid \mathcal{P} &\sim \pi(F_0, F_1 \mid \mathcal{P}) \\ \boldsymbol{X}_i \mid F_0, F_1, Z_i &\overset{\text{iid}}{\sim} (1 - Z_i)F_0 + Z_i F_1, \end{aligned} \tag{5.4}$$

for $i = 1, \ldots, n$, where $\pi(\mathcal{P})$ represents a prior on the space of all finite partitions, $\pi(F_0, F_1 \,|\, \mathcal{P})$ a joint prior for the two distributions (F_0, F_1) with densities such that $r = f_1/f_0$ is piecewise constant on $\mathcal{P}$.

The success of this covariate density modeling approach for inferring a balancing partition rests critically on the choice of the model on (F_0, F_1)—it must directly model the difference between the density ratio rather than modeling F_0 and F_1 separately and infer on r based on the joint posterior. We shall treat r as the key object of interest while the rest of F_0 and F_1 as the "nuisance parameters". The strategy we take to deal with them is through marginalization. It is thus desirable to construct a model that allow simple Rao-Blackwellization on the nuisance parameters and render direct marginal inference on the interesting "parameters" r and $\mathcal{P}$.

To complete (5.4), we need to specify $\pi(F_0, F_1 \,|\, \mathcal{P})$ and $\pi(\mathcal{P})$. Here we first present one Bayesian nonparametric model $\pi(F_0, F_1 \,|\, \mathcal{P})$ that satisfies the above considerations. We postpone the specification of $\pi(\mathcal{P})$ to Section 5.2.4 after introducing an

outcome model. Given $\mathcal{P}$, we specify (F_0, F_1) in terms of two components. The first component is the probability mass F_0 and F_1 have on each $A \in \mathcal{P}$, and the other is the conditional distribution $F_0(\cdot \mid A)$ and $F_1(\cdot \mid A)$ on each $A \in \mathcal{P}$. Now by the definition of $\mathcal{P}$, the two conditional distributions are identical on each $A \in \mathcal{P}$. Let $F_g(\mathcal{P}) = (F_g(A_1), F_g(A_2), \ldots, F_g(A_{|\mathcal{P}|}))$ for $g = 0, 1$. A simple model for the probability mass on $A \in \mathcal{P}$ is the Dirichlet prior:

$$F_g(\mathcal{P}) \overset{\text{ind}}{\sim} \text{Dirichlet}(b_1, b_2, \ldots, b_{|\mathcal{P}|}) \quad \text{for } g = 0, 1, \tag{5.5}$$

where $|\mathcal{P}|$ denotes the number of blocks in $\mathcal{P}$, $b_j = \nu G(A_j)$ for $\nu > 0$ and some base measure G. For the conditional distribution on each $A \in \mathcal{P}$, $F_0(\cdot \mid A) = F_1(\cdot \mid A)$, we shall adopt a flexible prior supported on all possible densities that can be easily integrated out. One class of such nonparametric processes is the so-called Pólya tree (PT) type models. For example, we can apply the OPT (Wong et al., 2010): $F_0(\cdot \mid A) = F_1(\cdot \mid A) = F_A(\cdot) \overset{\text{ind}}{\sim} \text{OPT}_{|A}(\mathcal{R}_A; \boldsymbol{\kappa}_A, \boldsymbol{\omega}_A, \rho_A)$ for all $A \in \mathcal{P}$. Putting the two pieces together, we shall denote our prior on (F_0, F_1) given $\mathcal{P}$ as

$$(F_0, F_1) \mid \mathcal{P} \sim \text{Dir-OPT}(\mathcal{P}; \nu, G; \mathcal{R}, \boldsymbol{\kappa}, \boldsymbol{\omega}, \rho), \tag{5.6}$$

where with a bit abuse of the notation, we use $\mathcal{R}, \boldsymbol{\kappa}, \boldsymbol{\omega}, \rho$ to denote the OPT parameters at any possible A.

Under this setup, the marginal likelihood of partition $\mathcal{P}$ can be written as

$$\begin{aligned} M(\mathcal{P}) &= p(\mathcal{X}_n \mid \mathcal{Z}_n, \mathcal{P}) \\ &= \int \prod_{i=1}^{n} f_0(\boldsymbol{x}_i)^{\mathbb{1}(z_i=0)} f_1(\boldsymbol{x}_i)^{\mathbb{1}(z_i=1)} d\,\pi(F_0, F_1 \mid \mathcal{P}) \\ &= M_0(\mathcal{P}) M_1(\mathcal{P}) \times \prod_{A \in \mathcal{P}} \Phi(A), \end{aligned} \tag{5.7}$$

where for $\mathcal{P} = \{A_1, A_2, \ldots, A_{|\mathcal{P}|}\}$, $g = 0, 1$, and $j = 1, \ldots, |\mathcal{P}|$,

$$M_g(\mathcal{P}) = \frac{\Gamma(\sum_j b_j)}{\prod_j \Gamma(b_j)} \frac{\prod_j \Gamma(b_j + n_i(A_j))}{\Gamma(\sum_j b_j + n_i)}, \tag{5.8}$$

which is the contribution to the marginal likelihood from the distribution of observations over the CDR partition in treatment group g, where $n_g(A)$ denotes the sample size in group g that fall in block A and n_g the total sample size in that group. $\Phi(A) = \int \prod_{\boldsymbol{x}_i \in A} f_A(\boldsymbol{x}_i \mid A) d\pi(F_A)$ is the marginal likelihood of the OPT process on each $A \in \mathcal{P}$, which summarizes the "nuisance" part of the model, with $f_A = dF_A/d\mu$ being the balanced covariate density in each block A. $\prod_{A \in \mathcal{P}} \Phi(A)$ is the contribution to the marginal likelihood from further details within each A, had the CDR be constant within each A. Intuitively, the empirical evidence will push further division over A if the decrease in the baseline contribution is small compared to the increase in $M_0(\mathcal{P})M_1(\mathcal{P})$ due to such a division.

A particularly desirable feature of the above model is its conjugacy. That is, given the data $\mathcal{D}_n$, the conditional posterior of (F_0, F_1) given $\mathcal{P}$ is of exactly the same form as the prior. In particular, *a posteriori*

$$F_g(\mathcal{P}) \,|\, \mathcal{P}, \mathcal{D} \overset{\text{ind}}{\sim} \text{Dirichlet}(b_1 + n_g(A_1), b_2 + n_g(A_2), \ldots, b_{|\mathcal{P}|} + n_g(A_{|\mathcal{P}|})),$$

for $g = 0, 1$. The conditional posterior for $F_A(\cdot) = F_0(\cdot \,|\, A) = F_1(\cdot \,|\, A)$ is also conjugate if PT type models are adopted (see for example Wong et al. (2010)), though since these are nuisance parameters we do not need them for our purpose. Our main interest lies in the conditional posterior of the CDRs

$$(r(A_1), r(A_2), \ldots, r(A_{|\mathcal{P}|})) = \left(\frac{F_1(A_1)}{F_0(A_1)}, \frac{F_1(A_2)}{F_0(A_2)}, \ldots, \frac{F_1(A_{|\mathcal{P}|})}{F_0(A_{|\mathcal{P}|})} \right), \tag{5.9}$$

which, *a posteriori* is the element-wise ratio of two independent Dirichlet random vectors.

5.2.3 Incorporating an outcome model

Heretofore, we construct a balancing partition $\mathcal{P}$ only with the conditional likelihood of $\mathcal{X}_n$ given $\mathcal{Z}_n$. In this section, we show that an outcome model can be incorporated

into the joint modeling framework in (5.4) for the heterogeneous treatment effect. Specifically, we specify a model for $\mathcal{Y}_n$ conditioning on $\mathcal{X}_n, \mathcal{Z}_n$ and $\mathcal{P}$.

Given $\mathcal{P}$, let $\mathcal{Y}_n(A), \mathcal{Z}_n(A)$ and $\mathcal{X}_n(A)$ be the vector of outcome, the vector of treatment assignment indicator and the covariate matrix of units in block $A \in \mathcal{P}$. As a starting point, suppose that the $\mathcal{Y}_n(A)$'s are conditionally independent for different $A \in \mathcal{P}$. For example, we can model $\mathcal{Y}_n(A)$ with a linear regression:

$$\begin{aligned} \mathcal{Y}_n(A) &= \mathbf{1}_{n(A)}\alpha(A) + \mathcal{Z}_n(A)\nu(A) + \mathcal{X}_n(A)\boldsymbol{\delta}(A) + \boldsymbol{\epsilon}(A) \\ \boldsymbol{\epsilon}(A) &\sim \mathrm{N}(\mathbf{0}_{n(A)}, \boldsymbol{I}_{n(A)}/w(A)), \end{aligned} \tag{5.10}$$

where $\mathbf{1}_{n(A)}, \mathbf{0}_{n(A)}$ are the $n(A)$ vector of ones and zeros, $\boldsymbol{I}_{n(A)}$ the identity matrix of size $n(A)$. To simplify the notations, we write the treatment indicator as the first dimension of the covariate matrix and use $\mathbf{x}(A) = (\mathcal{Z}_n(A), \mathcal{X}_n(A))$ to denote the treatment and the pre-treatment covariate as a whole. We also let $\boldsymbol{y}(A) = \mathcal{Y}_n(A)$. Moreover, we shall fix our attention on a specific block A and suppress the "(A)" and "$n(A)$" in the notations. We can then write the regression model at a specific block as

$$\boldsymbol{y} = \mathbf{1}\alpha + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}/w), \tag{5.11}$$

where $\boldsymbol{\beta} = (\nu, \boldsymbol{\delta}^\top)^\top$. We further assume that the covariates are centered: $\mathbf{1}^\top\mathbf{x} = \mathbf{0}$ and apply Zellner's g-prior (Zellner, 1986; Liang et al., 2008) on the parameters:

$$\pi(\alpha, w) \propto \frac{1}{w}, \quad \boldsymbol{\beta} \mid w \sim \mathrm{N}\left(\mathbf{0}, \frac{g}{w}(\mathbf{x}^\top\mathbf{x})^{-1}\right). \tag{5.12}$$

The marginal likelihood on A under this outcome model is available in closed form:

$$\begin{aligned} \Psi(A) &= \int (2\pi w^{-1})^{-n(A)/2} e^{-w(\boldsymbol{y}-\mathbf{1}\alpha-\mathbf{x}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{1}\alpha-\mathbf{x}\boldsymbol{\beta})/2} d\pi(\alpha, \boldsymbol{\beta}, w) \\ &= \frac{\Gamma((n-1)/2)}{\sqrt{\pi}^{(n-1)}\sqrt{n}} ||\boldsymbol{y} - \mathbf{1}\bar{y}||^{-(n-1)} \times \frac{(1+g)^{(n-2-p)/2}}{[1+g(1-R^2)]^{(n-1)/2}}, \end{aligned} \tag{5.13}$$

where R^2 is the ordinal coefficient of determination. Moreover, since the outcomes on each A are mutually independent, the conditional joint posterior of the regression parameters $(\boldsymbol{\beta}(A), w(A))$ are directly available through standard Bayesian updating:

$$\begin{aligned} w \mid \boldsymbol{y} &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2}\left[\frac{\hat{\boldsymbol{\beta}}^\top(\mathbf{x}^\top\mathbf{x})\hat{\boldsymbol{\beta}}}{(g+1)} + (\boldsymbol{y} - \mathbf{1}\bar{y} - \mathbf{x}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \mathbf{1}\bar{y} - \mathbf{x}\hat{\boldsymbol{\beta}})\right]\right) \\ \boldsymbol{\beta} \mid w, \boldsymbol{y} &\sim \text{N}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}, \frac{g}{g+1}(\mathbf{x}^\top\mathbf{x})^{-1}w^{-1}\right), \end{aligned} \tag{5.14}$$

where $\bar{y}$ is the mean of $\boldsymbol{y}$, $\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\boldsymbol{y}$ the MLE of $\boldsymbol{\beta}$.

Putting this outcome model together with the marginal likelihood from our covariate models, the full marginal likelihood given $\mathcal{P}$ becomes

$$\begin{aligned} M(\mathcal{P}) &= p(\mathcal{X}_n \mid \mathcal{Y}_n, \mathcal{Z}_n, \mathcal{P}) \\ &= p(\mathcal{X}_n \mid \mathcal{Z}_n, \mathcal{P}) \cdot p(\mathcal{Y}_n \mid \mathcal{X}_n, \mathcal{Z}_n, \mathcal{P}) \\ &= \left[M_0(\mathcal{P})M_1(\mathcal{P}) \prod_{A\in\mathcal{P}} \Phi(A)\right] \cdot \left[\prod_{A\in\mathcal{P}} \Psi(A)\right], \end{aligned} \tag{5.15}$$

where $\Psi(A)$ is the contribution to the marginal likelihood from the outcome model in $A \in \mathcal{P}$. Therefore, both the empirical evidence from the covariates and that from the outcome model will contribute to the marginal likelihood and inform the appropriate partition $\mathcal{P}$. In particular, the blocks in $\mathcal{P}$ will tend to be homogeneous both in the CDR (and thus the propensity score)*and* their outcome-treatment-covariate relationship. To inform the partitioning, only the marginal likelihood $\Psi(A)$ is needed. Therefore, as long as the marginal likelihood can be efficiently calculated, although not necessary, more complicated outcome models can also be used here.

It is worth noting that if one only uses the the contribution from the outcome model $\prod_{A\in\mathcal{P}} \Psi(A)$ is used as the marginal likelihood, this essentially boils down to fitting a Bayesian CART (Chipman et al., 1998) (in particular a treed piecewise

linear model), and the inferred $\mathcal{P}$ is aiming at achieving homogeneous treatment effects within each partition block. On the other hand, if only the contribution from the covariate model $M_0(\mathcal{P})M_1(\mathcal{P})\prod_{A\in\mathcal{P}}\Phi(A)$ is used as the marginal likelihood, the resulting model uncovers the design stage of the observational study and provides a matching strategy for causal comparison. These two ways of inferring $\mathcal{P}$ fall into a class of methods for modularized Bayesian inference (Liu et al., 2009), where factors of the likelihood function are ignored in the analysis. By incorporating both the treatment model *and* the outcome model into the marginal likelihood, the inferred partition $\mathcal{P}$ achieves homogeneity in both the CDR (or the propensity score) and the treatment effect. Therefore, this joint model achieves automatic covariate balancing through the partitioning $\mathcal{P}$ while fitting Bayesian CART, in particular a treed linear model, at the same time.

5.2.4 Priors on $\mathcal{P}$

To complete the specification of our probability model for $(\boldsymbol{X}_i, Y_i, \mathcal{P})$ given Z_i, we need to specify the prior on $\mathcal{P}$. In principle, any priors on finite partitions of the covariate space can be adopted, though different choices will result in different inference strategies. In this work we do not propose any new priors for partitions, but instead illustrate how common popular priors can be adopted. For example, a popular prior on the space of recursive partitions is the Bayesian CART prior (Chipman et al., 1998). Generally, such priors will result in marginal posteriors of $\mathcal{P}$ that are conjugate but with no closed-form expressions. These posteriors can be sampled from using MCMC algorithms.

These algorithms aim at sampling from the posterior $\pi(\mathcal{P}\,|\,\mathcal{D})$ through making local moves from $\mathcal{P}$ to a new partition $\mathcal{P}'$ on the space of finite partitions. Because each local move results in only a minor change in the partition, such as dividing one set into two halves, only a small portion of the contribution to the marginal

likelihood is changed under each move. Specifically, most of $\Phi(A)$ and $\Psi(A)$ remain the same, and one only needs to compute them for sets A that are exclusive to $\mathcal{P}$ or the proposal $\mathcal{P}'$. Let $\pi(\mathcal{P} \to \mathcal{P}')$ be the proposal kernel, then a Metropolis-Hastings sampler will accept the proposed new partition with probability

$$\min\left\{1, \frac{\pi(\mathcal{P}' \to \mathcal{P})\pi(\mathcal{P}')M(\mathcal{P}')}{\pi(\mathcal{P} \to \mathcal{P}')\pi(\mathcal{P})M(\mathcal{P})}\right\}. \tag{5.16}$$

Because $\mathcal{P}'$ and $\mathcal{P}$ differ only in a small number of A's, both $\pi(\mathcal{P}')/\pi(\mathcal{P})$ and $M(\mathcal{P}')/M(\mathcal{P})$ can be quickly computed at each step. Alternative strategies to MCMC include sequential Monte Carlo (SMC), as proposed by Lakshminarayanan et al. (2013). These algorithms are all directly applicable due to the available closed form marginal likelihood $M(\mathcal{P})$.

Recursive dimensionwise dyadic partitioning. We discuss a special version of the Bayesian CART prior, called recursive dimensionwise dyadic partitioning (RDDP), which leads to a particularly simple inference recipe through forward-backward recursion for calculating and sampling from the exact posterior without the need for approximate sampling algorithms such as MCMC and SMC when the number of covariates are not too large, e.g., (< 20). This is achieved through imposing that each set is allowed to be partitioned at the middle point of each continuous covariate. The middle point constraint substantially reduces the space of possible partitions, thereby allowing efficient computation.

Specifically, this prior on $\mathcal{P}$ can be described constructively as a generative procedure in the form of a coarse-to-fine recursive dyadic partitioning on the covariate space. We describe the procedure inductively. After k steps of partitioning, let $\mathcal{P}^{(k)}$ represent the current partition of $\Omega_{\boldsymbol{X}}$. Moreover, we write $\mathcal{P}^{(k)} = \mathcal{P}_0^{(k)} \cup \mathcal{P}_1^{(k)}$ where $\mathcal{P}_1^{(k)}$ are the partition blocks that can be partitioned further in future steps, whereas $\mathcal{P}_0^{(k)}$ are those blocks on which no further partition is done in future steps. (We shall

refer to blocks in $\mathcal{P}_0^{(k)}$ as the "stopped" blocks and those in $\mathcal{P}_1^{(k)}$ as "active" blocks. To begin the induction, we have $k = 0$, $\mathcal{P}^{(0)} = \mathcal{P}_1^{(0)} = \{\Omega_{\boldsymbol{X}}\}$, the trivial partition, and $\mathcal{P}_0^{(0)} = \varnothing$.

Now, the $(k+1)$-st step of the partitioning proceeds as follows. For each active block $A \in \mathcal{P}_1^{(k)}$, we draw a Bernoulli random variable

$$C(A) \sim \text{Bernoulli}(\gamma(A))$$

such that if $C(A) = 1$ then we do not further partition A, and move it from the active blocks into the stopped blocks. If $C(A) = 0$, then suppose there are $N(A)$ ways to divide A into two children (e.g., there are $N(A)$ covariates whose support is not atomic and each can be divided into halves). We randomly draw one of the $N(A)$ ways through generating a multinomial Bernoulli variable

$$D(A) \sim \text{Multi-Binom}(\boldsymbol{\lambda}(A)), \tag{5.17}$$

where $\boldsymbol{\lambda}(A) = (\lambda_1(A), \lambda_2(A), \ldots, \lambda_{N(A)}(A))$, and if $D(A) = d$ then divide A in the dth way into two children $A = A_l^d \cup A_r^d$. We then replace A with A_l^d and A_r^d in the active partition $\mathcal{P}_1^{(k+1)}$. This completes the inductive description of the $(k+1)$-st step of the partitioning.

One can show that as long as the stopping probability $\gamma(A) > \epsilon$ for all A and some $\epsilon > 0$, then this iterative partitioning will with probability 1 "stop" almost everywhere on $\Omega_{\boldsymbol{X}}$ (Wong et al., 2010). That is the "size" of the total size of $\mathcal{P}_1^{(k)}$ will decrease to zero as k increases. Practically, it is convenient to set a maximum number of steps in the partition, K, such that all sets are forced to stop if they are generated after K levels of partitioning. This completes the generative procedure for random partitions, and we say that the random partition arises from a RDDP prior with parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, and write it as

$$\mathcal{P} \sim \text{RDDP}(\mathcal{R}; \boldsymbol{\gamma}, \boldsymbol{\lambda}), \tag{5.18}$$

where $\mathcal{R}$ denotes the set of allowed partitions under the RDDP prior, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are the stopping indicators and the dimension selection indicators at any possible A. In this work, we shall focus on this RDDP prior on $\mathcal{P}$ for the simple and fast computation it offers.

5.3 Bayesian recursive partitioning for covariate balancing

In this section, we ignore the outcome information in constructing the partition $\mathcal{P}$ and focus on inferring a CDR partition through (5.4).

5.3.1 Posterior inference

To summarize, our prior on (F_0, F_1) takes the following hierarchical form

$$\begin{aligned} \mathcal{P} &\sim \text{RDDP}(\mathcal{R}; \boldsymbol{\gamma}, \boldsymbol{\lambda}) \\ (F_0, F_1) \mid \mathcal{P} &\sim \text{Dir-OPT}(\mathcal{P}; \nu, G; \mathcal{R}, \boldsymbol{\kappa}, \boldsymbol{\omega}, \rho), \end{aligned} \tag{5.19}$$

with an inductive description of this prior given in Prior 1. We now formally introduce the concept of force coupling. Let $K \in \mathbb{N}^+$ be the maximum level of the partition. That is, we force $F_0(\cdot \mid A) = F_1(\cdot \mid A) \overset{\text{ind}}{\sim} \text{OPT}_{|A}(\boldsymbol{\kappa}_A, \boldsymbol{\omega}_A, \rho_A)$ regardless of the value of $C(A)$ for $A \in \mathcal{P}_1^{(K)}$. We let $(F_0^{(K)}, F_1^{(K)})$ denote the covariate distributions generated from this process with maximum partition level K. Note that this inductive process naturally generates a partition tree of $\Omega_{\boldsymbol{X}}$. We denote this tree as $\mathcal{T}^{(K)}$. The set of internal nodes and leaves of $\mathcal{T}^{(K)}$ are denoted as $\mathcal{I}(\mathcal{T}^{(K)})$ and $\mathcal{L}(\mathcal{T}^{(K)})$, respectively. We have $\mathcal{L}(\mathcal{T}^{(K)}) = \mathcal{P}_0^{(K)}$ by definition. If the partition rule $\mathcal{R}$ allows only finite number of ways to partition each set, the total number of trees up to level K is finite. We denote this set of trees as $\mathscr{T}^{(K)}$. We also let $\mathscr{N}^{(K)} = \{A : A \in \bigcup_{\mathcal{T}^{(K)} \in \mathscr{T}^{(K)}} \mathcal{L}(\mathcal{T}^{(K)})\}$ to be all reachable subsets of $\Omega_{\boldsymbol{X}}$ by trees up to level K.

Prior 1 Two-step covariate modeling

procedure RDDP($\mathcal{R}; \boldsymbol{\gamma}, \boldsymbol{\lambda}$) ▷ **Step 1**: generate a partition
  Let $k = 0$, $\mathcal{P}_1^{(0)} = \{\Omega_{\boldsymbol{X}}\}$, $\mathcal{P}_0^{(0)} = \varnothing$, $\mathcal{P}^{(k)} = \mathcal{P}_0^{(k)} \cup \mathcal{P}_1^{(k)}$. ▷ Initialization
  while $\bigcup_{A \in \mathcal{P}_0^{(l)}} A \neq \Omega_{\boldsymbol{X}}$ **do** ▷ Inductively generate the partition
    $\mathcal{P}_0^{(k+1)} \leftarrow \mathcal{P}_0^{(k)}$, $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k)}$.
    for $A \in \mathcal{P}_1^{(k)}$ **do**
      Let $C(A) \sim$ Bernoulli($\gamma(A)$)
      if $C(A) = 1$ **then** ▷ Stopped
        $\mathcal{P}_0^{(k+1)} \leftarrow \mathcal{P}_0^{(k+1)} \cup \{A\}$.
        $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k+1)} - \{A\}$.
      end if
      if $C(A) = 0$ **then** ▷ Active
        Let $D(A) \sim$ Multi-Binom($\boldsymbol{\lambda}(A)$), $d \leftarrow D(A)$.
        $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k+1)} \cup \{A_l^d\} \cup \{A_r^d\} - \{A\}$.
      end if
    end for
    $l \leftarrow k + 1$. ▷ Move to next level
  end while
end procedure

$\mathcal{P} \leftarrow \mathcal{P}_0^{(k)}$ ▷ **Step 2:** covariate modeling given partition
Let $F_i(\mathcal{P}) \overset{\text{ind}}{\sim} \text{Dir}(b_1, b_2, \ldots, b_{|\mathcal{P}|})$, $i = 0, 1$.
for $A \in \mathcal{P}$ **do**
  Let $F_0(\cdot \mid A) = F_1(\cdot \mid A) \overset{\text{ind}}{\sim} \text{OPT}_{|A}(\boldsymbol{\kappa}_A, \boldsymbol{\omega}_A, \rho_A)$.
end for

return (F_0, F_1).

One advantage of the RDDP prior over other more flexible alternatives such as the Bayesian CART is that it facilitates simple and exact posterior inference. To see this, we first consider an alternative prior for (F_0, F_1) by modifying the RDDP generating process in the way shown in Prior 2. In this new formulation, the partition and the conditional model on the covariate are specified simultaneously. We denote this prior on the covariate distributions as

$$(F_0, F_1) \sim \text{RDDP-OPT}(\mathcal{R}; \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}; \boldsymbol{\kappa}, \boldsymbol{\omega}, \rho). \tag{5.20}$$

The following lemma shows that Prior 2 incorporates Prior 1 as a special case. There-

fore, posterior inference under Prior 1 can be achieved through posterior inference under Prior 2.

Lemma 1. *For $K \in \mathbb{N}^+$, suppose that $(F_0^{(K)}, F_1^{(K)})$ and $(\tilde{F}_0^{(K)}, \tilde{F}_1^{(K)})$ are generated by the procedure described in Prior 1 and Prior 2 with maximum partition level K respectively. We have $(\tilde{F}_0^{(K)}, \tilde{F}_1^{(K)}) \stackrel{d}{=} (F_0^{(K)}, F_1^{(K)})$ if $\alpha(A) = \nu G(A)$ for $A \in \mathscr{N}^{(K)}$.*

Prior 2 Simultaneous partitioning and covariate modeling

Let $k = 0$, $\mathcal{P}_1^{(0)} = \{\Omega_{\boldsymbol{X}}\}$, $\mathcal{P}_0^{(0)} = \varnothing$, $\mathcal{P}^{(k)} = \mathcal{P}_0^{(k)} \cup \mathcal{P}_1^{(k)}$. ▷ Initialization
while $\bigcup_{A \in \mathcal{P}_0^{(k)}} A \neq \Omega_{\boldsymbol{X}}$ **do** ▷ Inductively generate the partition
  $\mathcal{P}_0^{(k+1)} \leftarrow \mathcal{P}_0^{(k)}$, $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k)}$.
  for $A \in \mathcal{P}_1^{(k)}$ **do**
    Let $C(A) \sim \text{Bernoulli}(\gamma(A))$
    if $C(A) = 1$ **then** ▷ Stopped
      Let $F_0(\cdot \mid A) = F_1(\cdot \mid A) \overset{\text{ind}}{\sim} \text{OPT}_{\mid A}(\boldsymbol{\kappa}_A, \boldsymbol{\omega}_A, \rho_A)$.
      $\mathcal{P}_0^{(k+1)} \leftarrow \mathcal{P}_0^{(k+1)} \cup \{A\}$.
      $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k+1)} - \{A\}$.
    end if
    if $C(A) = 0$ **then** ▷ Active
      Let $D(A) \sim \text{Multi-Binom}(\boldsymbol{\lambda}(A))$, $d \leftarrow D(A)$.
      Let $F_i(A_l^d \mid A) \overset{\text{iid}}{\sim} \text{Beta}(\alpha(A_l^d), \alpha(A_r^d))$.
      $\mathcal{P}_1^{(k+1)} \leftarrow \mathcal{P}_1^{(k+1)} \cup \{A_l^d\} \cup \{A_r^d\} - \{A\}$.
    end if
  end for
  $k \leftarrow k + 1$. ▷ Move to next level
end while

return (F_0, F_1).

The next lemma shows that our model under Prior 2 is conjugate, and the posterior can be computed with a forward-backward algorithm. (The proofs of Lemma 1 and Lemma 2 are given in Appendix C.1.)

Lemma 2. *Suppose that $\boldsymbol{x}_0 = (\boldsymbol{x}_{01}, \boldsymbol{x}_{02}, \ldots, \boldsymbol{x}_{0n_0})$ and $\boldsymbol{x}_1 = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n_1})$ are i.i.d. samples from F_0 and F_1, where $(F_0, F_1) \sim$ RDDP-OPT$(\mathcal{R}; \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1; \boldsymbol{\kappa}, \boldsymbol{\omega}, \rho; K)$.*

The posterior of (F_0, F_1) *can be written as*

$$(F_0, F_1) \mid \boldsymbol{x}_0, \boldsymbol{x}_1 \sim \textit{RDDP-OPT}(\mathcal{R}; \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1; \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\omega}}, \tilde{\rho}; K), \tag{5.21}$$

where for each $A \in \mathscr{N}^{(K)}$*, we have*

1. *Stopping probabilities:*

$$\tilde{\gamma}(A) = \frac{\gamma(A)\phi(A)}{\eta(A)}.$$

2. *Partition selection probabilities: for* $d = 1, \ldots, N(A)$*,*

$$\tilde{\lambda}_d(A) = \lambda_d(A) \cdot \frac{B(n_0(A_l^d) + \alpha_0(A_l^d), n_0(A_r^d) + \alpha_0(A_r^d))}{B(\alpha_0(A_l^d), \alpha_0(A_r^d))}$$
$$\times \frac{B(n_1(A_l^d) + \alpha_1(A_l^d), n_1(A_r^d) + \alpha_1(A_r^d))}{B(\alpha_1(A_l^d), \alpha_1(A_r^d))} \times \eta(A_l^d) \cdot \eta(A_r^d) \times \frac{1 - \gamma(A)}{\eta(A) - \gamma(A)\phi(A)}.$$

3. *Probability assignment pseudo-counts: for* $i = 0, 1$*,*

$$\tilde{\alpha}_i(A) = n_i(A) + \alpha_i(A).$$

4. *OPT stopping probabilities:*

$$\tilde{\rho}(A) = \frac{\rho(A)u(A)}{\phi(A)}.$$

5. *OPT partition selection probabilities: for* $d = 1, \ldots, N(A)$

$$\tilde{\kappa}_d(A) = \kappa_d(A) \cdot \frac{B(n_1(A_l^d) + n_2(A_l^d) + \omega(A_l^d), n_1(A_r^d) + n_2(A_r^d) + \omega(A_r^d))}{B(\omega(A_l^d), \omega(A_r^d))}$$
$$\times \phi(A_l^d) \cdot \phi(A_r^d) \times \frac{1 - \rho(A)}{\phi(A) - \rho(A)u(A)}.$$

6. *OPT probability assignment pseudo-counts:*

$$\tilde{\omega}(A) = n_1(A) + n_2(A) + \omega(A).$$

where $n_i(A)$ denotes the number of sample from group i that fall into A,

$$\phi(A) = \rho(A)u(A) + (1-\rho(A))\sum_{d=1}^{p}\kappa_d(A)$$
$$\times \frac{B(n_1(A_l^d)+n_2(A_l^d)+\omega(A_l^d), n_1(A_r^d)+n_2(A_r^d)+\omega(A_r^d))}{B(\omega(A_l^d),\omega(A_r^d))} \times \phi(A_l^d)\cdot\phi(A_r^d),$$

$u(A)$ is the conditional likelihood of the base distribution, and

$$\eta(A) = \gamma(A)\phi(A) + (1-\gamma(A))\sum_{d=1}^{p}\lambda_d(A)\frac{B(n_1(A_l^d)+\alpha_1(A_l^d), n_1(A_r^d)+\alpha_1(A_r^d))}{B(\alpha_1(A_l^d),\alpha_1(A_r^d))}$$
$$\times \frac{B(n_2(A_l^d)+\alpha_2(A_l^d), n_2(A_r^d)+\alpha_2(A_r^d))}{B(\alpha_2(A_l^d),\alpha_2(A_r^d))}\cdot\eta(A_l^d)\cdot\eta(A_r^d).$$

if $l(A) < K$ and $\eta(A) = \phi(A)$ if $l(A) = K$, where $l(A)$ denotes the level of A.

5.3.2 Estimation of the average treatment effect

In this section, we consider estimating population treatment effects such as the average treatment effect (ATE): $\tau^{\text{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}_{\boldsymbol{X}}[\tau(\boldsymbol{X})]$. Suppose that we have i.i.d. samples $(\boldsymbol{X}_1, Z_1, Y_1), (\boldsymbol{X}_2, Z_2, Y_2), \ldots, (\boldsymbol{X}_n, Z_n, Y_n)$. Given a balancing partition $\mathcal{P}$, one estimator of τ^{ATE} is (Lunceford and Davidian, 2004)

$$\hat{\tau}_{\mathcal{P}} = \sum_{A\in\mathcal{P}}\frac{n(A)}{n}\left(n_1(A)^{-1}\sum_{i=1}^{n}Z_iY_i\mathbb{1}_{(\boldsymbol{X}_i\in A)} - n_0(A)^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i\mathbb{1}_{(\boldsymbol{X}_i\in A)}\right). \quad (5.22)$$

We define the posterior average of $\hat{\tau}_{\mathcal{P}}$ with respect to $\Pi(\mathcal{P} \mid \mathcal{X}_n)$ as

$$\mathbb{E}[\hat{\tau}_{\mathcal{P}} \mid \mathcal{X}_n, \mathcal{Z}_n] = \int \hat{\tau}_{\mathcal{P}} d\Pi(\mathcal{P} \mid \mathcal{X}_n, \mathcal{Z}_n). \quad (5.23)$$

In practice, this expectation can be estimated by drawing S partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_S$ from $\Pi(\mathcal{P} \mid \mathcal{X}_n, \mathcal{Z}_n)$, and take the average of $\hat{\tau}_{\mathcal{P}_s}$. The following theorems establish the consistency of this estimator for estimating the ATE:

Theorem 1. *Suppose that $\mathcal{P}$ is generated through the RDDP prior with maximum level K for $K \geqslant 0$. Suppose that the following assumptions hold:*

1. *As $n \to \infty$, $n_0/(n_0+n_1) \to \zeta_0 \in (0,1)$, $n_1/(n_0+n_1) \to \zeta_1 \in (0,1)$, where n_0 and n_1 are the number of control and treatment unit, respectively;*
2. *The CDR $r(\boldsymbol{x})$ is uniformly continuous;*
3. *The potential outcomes are bounded such that $|Y_i(0)|, |Y_i(1)| \leqslant B$;*
4. *The covariate densities are bounded and bounded away from zero: $0 < \delta < f_0(\boldsymbol{x}), f_1(\boldsymbol{x}) < M$.*

Then

$$\mathbb{E}[\hat{\tau}_{\mathcal{P}} \mid \mathcal{X}_n, \mathcal{Z}_n] \xrightarrow{p} \tau^{\text{ATE}},$$

as $n \to \infty$ and $K \to \infty$.

Theorem 2. *Suppose that $\mathcal{P}$ is generated through the RDDP prior with maximum level K for $K \geqslant 0$. Under the assumption of Theorem 1 and suppose that in addition, $r(\boldsymbol{x})$ is Lipschitz continuous with a Lipschitz constant L_r. Then*

$$|\mathbb{E}[\hat{\tau}_{\mathcal{P}} \mid \mathcal{X}_n, \mathcal{Y}_n] - \tau^{\text{ATE}}| < (C_0 + C_1) \cdot \left(\frac{M^5 \bar{r}}{\delta^5 \underline{r}} \cdot L_r^2 \cdot \frac{p}{4^{\lfloor K/p \rfloor}} \right)^{\frac{1}{3}}.$$

as $n \to \infty$, where

$$C_0 = B \left[\frac{2}{(1-\bar{e})^2} + M + \frac{M}{1-\bar{e}} + \frac{\zeta_1 \bar{r} M}{\zeta_0} \right]$$

$$C_1 = B \left[\frac{2}{\underline{e}^2} + M + \frac{M}{\underline{e}} + \frac{\zeta_0 M}{\zeta_1 \underline{r}} \right],$$

$\underline{r}$ and $\bar{r}$ are the induced bounds on the CDR such that $0 < \underline{r} \leqslant r(\boldsymbol{x}) \leqslant \bar{r}$, $\underline{e}$ and $\bar{e}$ are the induced bounds on the propensity scores such that $0 < \underline{e} \leqslant e(\boldsymbol{x}) \leqslant \bar{e}$.

The proofs of Theorem 1 and Theorem 2 are given in Appendix C.3. We provide additional details on theoretical properties of the posterior of our CDR model in Appendix C.2.

In practice, this estimator is always biased due to remaining imbalance in each block. To alleviate this issue, model-based adjustment can be applied within each block. Suppose that

$$Y_i = \sum_{j=1}^{|\mathcal{P}|} m(Z_i, \boldsymbol{X}_i, \boldsymbol{a}_j)\mathbb{1}_{(\boldsymbol{X}_i \in A_j)} + \epsilon_i, \quad \epsilon_i \overset{\text{ind}}{\sim} \mathrm{N}\left(0, \sum_{j=1}^{K} s_j^2 \mathbb{1}_{(\boldsymbol{X}_i \in A_j)}\right), \tag{5.24}$$

where $m(Z, \boldsymbol{X}, \boldsymbol{a}_j)$ represents a model on each block with the block-specific parameter $\boldsymbol{a}_j$, $\hat{\boldsymbol{a}}_j$ the estimate of $\boldsymbol{a}_j$, for $j = 1, 2, \ldots, |\mathcal{P}|$. This leads to the following model-adjusted variant of $\hat{\tau}_{\mathcal{P}}$ (Lunceford and Davidian, 2004):

$$\hat{\tau}_{\mathcal{P}}^{(m)} = \sum_{j=1}^{|\mathcal{P}|} \frac{n(A_j)}{n} \left(n_1(A_j)^{-1} \sum_{i=1}^{n} \mathbb{1}_{(\boldsymbol{X}_i \in A_j)} [m(1, \boldsymbol{X}_i, \hat{\boldsymbol{a}}_j) - m(0, \boldsymbol{X}_i, \hat{\boldsymbol{a}}_j)] \right). \tag{5.25}$$

For example, $m(Z_i, \boldsymbol{X}_i, \boldsymbol{a}_j) = a_{0j} + a_{Zj} Z_i + \boldsymbol{X}_i^\top \boldsymbol{a}_{\boldsymbol{X}j}$ corresponds to the linear regression adjustment. When $\boldsymbol{a}_{\boldsymbol{X}j} = \boldsymbol{0}$, $\hat{\tau}_{\mathcal{P}}^{(m)}$ degenerate to $\hat{\tau}_{\mathcal{P}}$. Since the block-specific model adjustment are independent, an estimator of the variance of $\hat{\tau}_{\mathcal{P}}^{(m)}$ is

$$\hat{\sigma}_{\mathcal{P}}^{2(m)} = \sum_{j=1}^{|\mathcal{P}|} \left(\frac{n(A_j)}{n} \right)^2 \hat{\sigma}_j^{2(m)}, \tag{5.26}$$

where $\hat{\sigma}_j^{2(m)}$ is an estimation of the variance of $[m(1, \boldsymbol{X}_i, \hat{\boldsymbol{a}}_j) - m(0, \boldsymbol{X}_i, \hat{\boldsymbol{a}}_j)]$ for block j. For the linear regression model, $\hat{\sigma}_j^{2(m)}$ is the sampling variance of α_{jZ}.

Note that (5.26) estimates the variance of $\hat{\tau}_{\mathcal{P}}^{(m)}$ for a given partition, ignoring the uncertainties in constructing $\mathcal{P}$. Recall that $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_S$ are independent draws from $\Pi(\mathcal{P} \mid \mathcal{X}_n, \mathcal{Z}_n)$. Let

$$\hat{\tau}^{(m)} = \frac{1}{S} \sum_{s=1}^{S} \hat{\tau}_{\mathcal{P}_s}^{(m)}. \tag{5.27}$$

We shall use $\hat{\tau}^{(m)}$ as an estimation to the ATE. This estimator has a close connection to multiple imputation (MI) of missing data. In our case, the underlying design is "missing". The S partitions drawn from the posterior can be viewed as the multiple imputations to this missing design. Following this discussion, the variance of $\hat{\tau}^{(m)}$ can be written as

$$\mathbb{V}[\hat{\tau}_{\mathcal{P}}^{(m)} \mid \mathcal{X}_n, \mathcal{Z}_n] = \mathbb{E}(\mathbb{V}[\hat{\tau}_{\mathcal{P}}^{(m)} \mid \mathcal{X}_n, \mathcal{Z}_n, \mathcal{P}]) + \mathbb{V}(\mathbb{E}[\hat{\tau}_{\mathcal{P}}^{(m)} \mid \mathcal{X}_n, \mathcal{Z}_n, \mathcal{P}]). \tag{5.28}$$

and can be estimated with

$$\hat{\sigma}^{2(m)} \approx \frac{1}{S}\sum_{s=1}^{S} \hat{\sigma}_{\mathcal{P}_s}^{2(m)} + \left(1 + \frac{1}{S}\right) \frac{\sum_{s=1}^{S} (\hat{\tau}_{\mathcal{P}_s}^{(m)} - \hat{\tau}^{(m)})^2}{S-1}. \tag{5.29}$$

A similar strategy has been discussed in Liao and Zigler (2020) to incorporate uncertainties in estimating the propensity score for the propensity score based methods.

Incorporating the propensity score. A major limitation of the current RDDP prior and the Bayesian CART prior in general is that all the splits are perpendicular to some covariate axis. Similar to the variable combination procedures used to strengthen a CART model (Breiman et al., 1984), the covariates in the RDDP model can be combined to provide other possible split directions. In the causal setting, a natural split direction is given by the estimated propensity score. Let $\hat{e}(\boldsymbol{X}_i)$ be an estimated propensity score for $i = 1, 2, \ldots, n$. We can treat $\hat{e}(\boldsymbol{X}_i)$ as another dimension of the pretreatment covariate $\boldsymbol{X}_i$. The specific model used to estimate the propensity score is not a concern here since this extra direction is only served as a possible shortcut instead of the only way towards covariate balancing. One can even incorporate estimated propensity scores from different models to offer different shortcuts.

By allowing split in the propensity score direction, our retrospective matching method can be viewed as a direct extension to the classic propensity score stratification (Rosenbaum and Rubin, 1984). If only the propensity score dimension is split in

$\mathcal{P}$, our model essentially performs propensity score stratification that automatically determines the splitting points and the number of strata. If splits also happen in the original covariate space, it suggests that utilizing information in the covariates that are not synthesized in the estimated propensity score offers a better chance to achieve the in-sample covariate balancing.

This viewpoint is especially useful when the number of covariates is large (e.g., > 20) such that exact inference under the RDDP prior with all the covariates is practically infeasible. In this case, if some good estimates of the propensity score is available, one can let the covariate space to include the estimated propensity score and only a subset of important covariates. Now the propensity score model is essential since it is the only bridge towards the covariate balancing.

5.3.3 Numerical examples

In this section, we replicate the classic simulation experiment in Kang et al. (2007), which was originally proposed to evaluate the performance of the doubly robust estimator (Robins and Rotnitzky, 1995).

Data generating process: For $i = 1, \ldots, n$, suppose that there are four pretreatment covariates $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$, $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \Sigma)$, $\boldsymbol{\Sigma} = \mathrm{diag}(1, 1, 1, 1)$. Consider the following true propensity score model and the true outcome model:

$$\begin{aligned} \mathrm{logit}(e_i(\boldsymbol{x}_i)) &= -x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4} \\ y_i &= 210 + 27.4x_{i1} + 13.7x_{i2} + 13.7x_{i3} + 13.7x_{i4} + \epsilon_i, \end{aligned} \tag{5.30}$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$. In this setup, a logistic regression of the covariates would be a correctly specified propensity score model, and a linear regression of y_i on $\boldsymbol{x}_i$ would be the correct outcome model.

Competitors: We compare our method (referred to as RP) with the following meth-

ods: the causal random forest (CF), matching based on the Mahalanobis distance (MC), and three propensity score based methods: the Horvitz-Thompson estimator (HT), the inverse propensity score weighting estimator (IPW) and the doubly robust estimator (DR):

$$\begin{aligned}
\hat{\tau}_{\mathrm{HT}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \\
\hat{\tau}_{\mathrm{IPW}} &= \frac{\sum_{i=1}^{n} Z_i Y_1/\hat{e}_i}{\sum_{i=1}^{n} Z_i/\hat{e}_i} - \frac{\sum_{i=1}^{n}(1-Z_i)Y_1/(1-\hat{e}_i)}{\sum_{i=1}^{n}(1-Z_i)/(1-\hat{e}_i)} \\
\hat{\tau}_{\mathrm{DR}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i-\hat{e}_i)\hat{m}_1(\boldsymbol{X}_i)}{\hat{e}_i}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-\hat{e}_i} + \frac{(Z_i-\hat{e}_i)\hat{m}_0(\boldsymbol{X}_i)}{1-\hat{e}_i}\right],
\end{aligned} \tag{5.31}$$

where $\hat{m}_1(\boldsymbol{X}_i)$ and $\hat{m}_0(\boldsymbol{X}_i)$ are the OLS estimates of Y_i under the linear model given $Z_i = 1$ or 0. For the propensity score based methods, we estimate the propensity score through two parametric methods: the standard logistic regression and the CBPS (Imai and Ratkovic, 2014). Note that RP, CF and MC are essentially modeling the propensity score nonparametrically. RP and DR also explicitly make use of a parametric outcome model. CF models the outcome in a nonparametric way.

Experiments: As in Kang et al. (2007), we consider four scenarios. In scenario I and II, we suppose that the true covariates $\boldsymbol{x}_i$'s are observed directly. In scenario III and IV, instead of being given $\boldsymbol{x}_i$'s directly, we assume that the researcher only observes $\boldsymbol{w}_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})$, which is a nonlinear transform of the original covariates:

$$\begin{aligned}
w_{i1} &= \exp(x_{i1}/2) \\
w_{i2} &= x_{i2}/(1+\exp(x_{i1})) + 10 \\
w_{i3} &= (x_{i1}x_{i3}/25 + 0.6)^3 \\
w_{i4} &= (x_{i2} + x_{i4} + 20)^2.
\end{aligned} \tag{5.32}$$

Table 5.1: RMSE for the competing methods.

n	RP	CF	MC	HT	IPW	DR	HT	IPW	DR
				CBPS			GLM		
I									
200	0.17	2.89	5.47	12.41	4.29	0.16	17.29	3.63	0.18
600	0.10	1.28	3.66	7.42	2.53	0.10	8.42	1.96	0.10
1500	0.06	0.68	2.64	4.84	1.52	0.06	5.20	1.22	0.06
II									
200	2.62	2.89	5.47	12.41	4.29	2.11	17.29	3.63	2.66
600	1.48	1.28	3.66	7.42	2.53	1.34	8.42	1.96	1.48
1500	1.17	0.67	2.64	4.84	1.52	0.91	5.20	1.22	0.94
III									
200	0.17	5.28	7.35	13.61	8.22	0.17	108.98	8.74	0.42
600	0.09	2.85	4.95	8.99	6.76	0.10	206.47	9.87	0.48
1500	0.06	1.73	3.62	8.27	5.86	0.07	1738.94	11.17	4.44
IV									
200	3.91	5.28	7.35	13.61	8.22	6.05	108.98	8.74	16.43
600	3.16	2.85	4.95	8.99	6.76	6.10	206.47	9.87	70.64
1500	3.12	1.73	3.62	8.27	5.86	6.64	1738.94	11.17	957.07

In this case, the logistic regression propensity score model and the linear regression outcome model are misspecified. In scenario I and III, we assume that the outcome model for RP and DR are correctly specified, regardless of the covariates observed. We consider three sample size: 200, 600 and 1500. We randomly simulate 5000 datasets for each sample size and apply each competing method in the four scenarios to estimate the ATE (the true ATE is zero in all scenarios). The estimated Root Mean Square Errors (RMSE) are shown in Table 5.1. (For CF, we use the `average_treatment_effect` function in the `R` package `grf`. For MC, we use the `Match` function in the `R` package `Matching`. For CBPS, we use the `CBPS` function in the `R` package `CBPS`.)

Analyses: Table 5.1 shows that the overall performance of RP is promising. With

the induced nonparametric propensity score model, RP is capable of balancing the covariates when the transformed covariates are observed. With a correctly specified outcome model, the RMSE of RP reduces significantly, indicating its doubly-robust nature. We note that as the sample size increases, the performance of RP improves slower comparing to the other nonparametric methods CF and MC. This is mainly due to the restricted partitioning scheme under the RDDP prior.

6

Conclusions

In this work, we presented three novel Bayesian hierarchical models. In Chapter 3, we introduced BGCR for solving the two-sample problem in the microbiome context. BGCR transforms the two-sample problem into a multiple testing problem and improves the power of these tests with a graphical model. BGCR also allows controlling for confounding covariates and provides a flexible way to report and visualize the testing results. In Chapter 4, we proposed DTMM as a tool for clustering microbiome count data in a model-based way. DTMM simultaneously selects relevant nodes in the phylogenetic tree and performs clustering, which is of scientific and statistical importance. In Chapter 5, we introduced a general framework for causal inference. Our framework allows two ways to estimate the treatment effect, either through the traditional covariate balancing strategy or by joint modeling the covariates and the outcome.

As we saw, an underlying tree-structure plays a crucial role in each of these methods. For BGCR and DTMM, a phylogenetic tree is incorporated to construct a more flexible generative model for microbiome count data than the well-known Dirichlet-multinomial model while keeping inference tractable. The phylogenetic

tree essentially imposes a set of modeling assumptions on the covariate structure among OTUs—that this structure can be well represented by the tree. In fact, the formulating of BGCR and DTMM requires only a fixed tree over the OTUs, not necessarily the phylogenetic tree. There can be trees that are better than the phylogenetic tree in terms of characterizing the dependency structure among OTUs. Searching for such trees and evaluating the benefits of using them instead of the phylogenetic tree in BGCR and DTMM is an important direction to explore. For our causal model, the tree is induced by the recursive partitioning scheme. Unlike in the microbiome setting, this tree is treated as a parameter and is learned from the data. In this setting, we are interested not in the tree itself, but the resulting balancing partition of the covariate space represented by its leaves.

These three methods are closely related to the Pòlya tree type models in the Bayesian nonparametric literature. For example, Ma and Wong (2011) and Soriano and Ma (2017) adopt tree-based models to the general two-sample problem; Christensen and Ma (2020) achieves clustering functional data through the hierarchical adaptive Pòlya tree. Although motivated by features of the microbiome data, some novel developments of BGCR and DTMM can be adopted in more general settings. For example, the regression adjustment in BGCR can be adopted in the MRS test (Soriano and Ma, 2017) to control for false discoveries when covariate information is available. Similarly, the node selection module of DTMM can be applied to the more general functional data clustering problem. Our retrospective framework for causal inference is closely related to the classic two-sample problem. By conditioning on the treatment assignment indicator, we are essentially comparing the covariate distributions under the two treatment groups. The theories we developed here are thus applicable to the coupling optional Pòlya tree (Ma and Wong, 2011) for the two-sample problem.

Appendix A

Supplementary Material for Chapter 3

A.1 Computational strategies

The crux of doing inference with BCR is to calculate the marginal likelihoods $M_s(A)$. We use the computational strategies in Ma and Soriano (2018) to compute the integrals. Specifically, for fixed ν, the inner integrals on the regression parameters are evaluated based on Laplace approximation; then the outer integral on ν is calculated with finite Riemann approximation.

With a bit abuse of the notations, in this section, we use $\boldsymbol{\beta}(A)$ to denote the 'active' regression coefficient in (3.10). That is, under the null, $\boldsymbol{\beta}(A)$ is just the $\boldsymbol{\beta}(A)$ in (3.10); under the alternative, $\boldsymbol{\beta}(A)$ denotes $(\boldsymbol{\beta}(A)^\top, \gamma(A))^\top$. $\boldsymbol{x}_{ij}$ is redefined to be the 'active' covariates in the same sense. With these notations, we have $g(\theta_{\boldsymbol{x}_{ij}}(A)) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}(A)$ and $\theta_{ij}(A) \mid \boldsymbol{x}_{ij}, \nu(A) \sim \text{Beta}(\theta_{\boldsymbol{x}_{ij}}(A)\nu(A), (1-\theta_{\boldsymbol{x}_{ij}}(A))\nu(A))$ for each local beta-binomial model.

The computational strategies are the same under both hypotheses. Let $\pi_A(\boldsymbol{\beta})$ be the prior density of $\boldsymbol{\beta}(A)$ under either hypothesis. For fixed ν in the support of

$G_A(\nu)$, by the Laplace approximation, the inner integral is

$$\begin{aligned}
L_\nu(A) &= \int \prod_{i=1}^{2} \prod_{j=1}^{n_i} \mathcal{L}_{BB}(g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}), \nu \mid y_{ij}(A_l), y_{ij}(A_r)) \pi_A(\boldsymbol{\beta}) d\boldsymbol{\beta} \\
&= \int \exp\left\{ \sum_{i,j} \log \mathcal{L}_{BB}(g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}), \nu \mid y_{ij}(A_l), y_{ij}(A_r)) + \log \pi_A(\boldsymbol{\beta}) \right\} d\boldsymbol{\beta} \\
&= \int \exp\left\{ h_\nu(\boldsymbol{\beta}) \right\} d\boldsymbol{\beta} \\
&\approx \exp\left\{ h_\nu(\hat{\boldsymbol{\beta}}_\nu) \right\} \cdot (2\pi)^{d/2} \cdot | - H_\nu(\hat{\boldsymbol{\beta}}_\nu)|^{-1/2} \\
&= \hat{L}_\nu(A)
\end{aligned}$$

where $h_\nu(\boldsymbol{\beta}) = \sum_{i,j} \log \mathcal{L}_{BB}(g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}), \nu \mid y_{ij}(A_l), y_{ij}(A_r)) + \log \pi_A(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_\nu$ is the maximizer of $h_\nu(\boldsymbol{\beta})$, $H_\nu(\hat{\boldsymbol{\beta}}_\nu)$ is the Hessian matrix of $h_\nu(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_\nu$. d is the degrees of freedom of $\boldsymbol{\beta}$, which is $(p+2)$ under the alternative and $(p+1)$ under the null. We describe a Newton-Raphson algorithm to solve for $\hat{\boldsymbol{\beta}}_\nu$ below. The log-likelihood function is strictly log-concave and the Newton-Raphson method generally converges after only a few iterations. Finally, to get $M_s(A)$, we compute the outer integral on ν, $\int L_\nu(A) dG_A(\nu)$, with finite Riemann approximations. Specifically, after calculating $\hat{L}_\nu(A)$ at a grid of ν's: $\nu_1, \nu_2, \ldots, \nu_M$, we have

$$\int L_\nu(A) dG_A(\nu) \approx \sum_{m=2}^{M} \hat{L}_{\nu_m}(A)(G_A(\nu_m) - G_A(\nu_{m-1})).$$

Newton-Rhaphson for $\hat{\boldsymbol{\beta}}_\nu$. In this subsection, we shall fix our attention on a specific node A and suppress the '(A)' in the notations. Moreover, we let $y_{ij}(A_l) = t_{ij}$ for simplicity and express the local Beta-Binomial regression model on A as

$$t_{ij} \sim \text{Binomial}(y_{ij}, \theta_{ij}), \quad \theta_{ij} \sim \text{Beta}(\theta_{\boldsymbol{x}_{ij}}\nu, (1-\theta_{\boldsymbol{x}_{ij}})\nu), \quad \text{and} \quad g(\theta_{\boldsymbol{x}}) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

The contribution to the log marginal likelihood from the j-th observation in group i is

$$\begin{aligned} l_{ij} &= \log \mathcal{L}_{BB}(g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}), \nu \mid t_{ij}, y_{ij} - t_{ij}) \\ &= \log \Gamma(\theta_{\boldsymbol{x}_{ij}} \nu + t_{ij}) + \log \Gamma((1 - \theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \log \Gamma(\nu + y_{ij}) \\ &\quad - \log \Gamma(\theta_{\boldsymbol{x}_{ij}} \nu) - \log \Gamma((1 - \theta_{\boldsymbol{x}_{ij}})\nu) + \log \Gamma(\nu). \end{aligned}$$

Taking the first derivative w.r.t. $\boldsymbol{\beta}$,

$$\frac{\partial l_{ij}}{\partial \boldsymbol{\beta}} = \frac{\partial l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}} \cdot \frac{\partial \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}} \cdot \frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}$$

where $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Now with ϕ denoting the digamma function,

$$\frac{\partial l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}} = \nu[\phi(\theta_{\boldsymbol{x}_{ij}} \nu + t_{ij}) - \phi((1 - \theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \phi(\theta_{\boldsymbol{x}_{ij}} \nu) + \phi((1 - \theta_{\boldsymbol{x}_{ij}})\nu)].$$

With the logit link, $\theta_{\boldsymbol{x}_{ij}} = g^{-1}(\eta_{ij}) = 1/(1 + e^{-\eta_{ij}})$, and

$$\frac{\partial \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}} = (g^{-1})'(\eta_{ij}) = \theta_{\boldsymbol{x}_{ij}}(1 - \theta_{\boldsymbol{x}_{ij}}).$$

Thus

$$\begin{aligned} \frac{\partial l_{ij}}{\partial \boldsymbol{\beta}} &= \nu \theta_{\boldsymbol{x}_{ij}}(1 - \theta_{\boldsymbol{x}_{ij}})[\phi(\theta_{\boldsymbol{x}_{ij}} \nu + t_{ij}) - \phi((1 - \theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) \\ &\quad - \phi(\theta_{\boldsymbol{x}_{ij}} \nu) + \phi((1 - \theta_{\boldsymbol{x}_{ij}})\nu)]\boldsymbol{x}_{ij}. \end{aligned}$$

The second derivative of l_{ij} w.r.t. $\boldsymbol{\beta}$ is

$$\begin{aligned} \frac{\partial^2 l_{ij}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \frac{\partial^2 l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}^2} \cdot \left(\frac{\partial \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}}\right)^2 \cdot \left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}^\top}\right) + \frac{\partial l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}} \cdot \frac{\partial^2 \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}^2} \cdot \left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}^\top}\right) \\ &\quad + \frac{\partial l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}} \cdot \frac{\partial \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}} \frac{\partial^2 \eta_{ij}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}. \end{aligned}$$

The third term on the right-hand side is equal to zero. With ψ being the trigamma function,

$$\frac{\partial^2 l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}^2} = \nu^2[\psi(\theta_{\boldsymbol{x}_{ij}} \nu + t_{ij}) + \psi((1 - \theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \psi(\theta_{\boldsymbol{x}_{ij}} \nu) - \psi((1 - \theta_{\boldsymbol{x}_{ij}})\nu)].$$

Thus the first term is

$$\frac{\partial^2 l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}^2} \cdot \left(\frac{\partial \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}}\right)^2 \cdot \left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)^\top$$

$$= \nu^2[\psi(\theta_{\boldsymbol{x}_{ij}}\nu + t_{ij}) + \psi((1-\theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij})$$

$$- \psi(\theta_{\boldsymbol{x}_{ij}}\nu) - \psi((1-\theta_{\boldsymbol{x}_{ij}})\nu)]\theta_{\boldsymbol{x}_{ij}}^2(1-\theta_{\boldsymbol{x}_{ij}})^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top.$$

The second term, which has expectation zero, is

$$\frac{\partial l_{ij}}{\partial \theta_{\boldsymbol{x}_{ij}}} \cdot \frac{\partial^2 \theta_{\boldsymbol{x}_{ij}}}{\partial \eta_{ij}^2} \cdot \left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}}\right)^\top$$

$$= \nu\left[\phi(\theta_{\boldsymbol{x}_{ij}}\nu + t_{ij}) - \phi((1-\theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \phi(\theta_{\boldsymbol{x}_{ij}}\nu) + \phi((1-\theta_{\boldsymbol{x}_{ij}})\nu)\right]$$

$$\times \theta_{\boldsymbol{x}_{ij}}(1-\theta_{\boldsymbol{x}_{ij}})(1-2\theta_{\boldsymbol{x}_{ij}})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top.$$

For each $i = 1, 2$, $j = 1, 2, \ldots n_i$, let

$$a_{ij} = \phi(\theta_{\boldsymbol{x}_{ij}}\nu + t_{ij}) - \phi((1-\theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \phi(\theta_{\boldsymbol{x}_{ij}}\nu) + \phi((1-\theta_{\boldsymbol{x}_{ij}})\nu)$$

$$b_{ij} = \psi(\theta_{\boldsymbol{x}_{ij}}\nu + t_{ij}) + \psi((1-\theta_{\boldsymbol{x}_{ij}})\nu + y_{ij} - t_{ij}) - \psi(\theta_{\boldsymbol{x}_{ij}}\nu) - \psi((1-\theta_{\boldsymbol{x}_{ij}})\nu).$$

Since the total log likelihood is $l = \sum_{i,j} l_{ij}$,

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \nu \sum_{i,j} a_{ij}\theta_{\boldsymbol{x}_{ij}}(1-\theta_{\boldsymbol{x}_{ij}})\boldsymbol{x}_{ij} = \nu \boldsymbol{X}^\top W_1 \boldsymbol{z},$$

where the rows of $\boldsymbol{X}$ are $\boldsymbol{x}_{ij}^\top$, $W_1 = \text{diag}(a_{ij})$ and $\boldsymbol{z} = (\theta_{\boldsymbol{x}_{11}}(1-\theta_{\boldsymbol{x}_{11}}), \ldots, \theta_{\boldsymbol{x}_{1n_1}}(1-\theta_{\boldsymbol{x}_{1n_1}}), \ldots, \theta_{\boldsymbol{x}_{2n_2}}(1-\theta_{\boldsymbol{x}_{2n_2}}))^\top$. The rows of $\boldsymbol{X}, W_1$ and the elements of $\boldsymbol{z}$ are ordered first by j and then i.

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = \nu^2 \sum_{i,j} b_{ij}\theta_{\boldsymbol{x}_{ij}}^2(1-\theta_{\boldsymbol{x}_{ij}})^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top + \nu \sum_{i,j} a_{ij}\theta_{\boldsymbol{x}_{ij}}(1-\theta_{\boldsymbol{x}_{ij}})(1-2\theta_{\boldsymbol{x}_{ij}})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top$$

$$= -\nu \boldsymbol{X}^\top W_2 \boldsymbol{X},$$

where $W_2 = -\text{diag}(\nu b_{ij}\theta^2_{\boldsymbol{x}_{ij}}(1-\theta_{\boldsymbol{x}_{ij}})^2 + a_{ij}\theta_{\boldsymbol{x}_{ij}}(1-\theta_{\boldsymbol{x}_{ij}})(1-2\theta_{\boldsymbol{x}_{ij}}))$. The columns of W_2 is also ordered first by j and then by i.

When applying Laplace approximation to evaluate the marginal likelihood for a fixed ν,

$$L_\nu = \int \exp\{l(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta})\} d\boldsymbol{\beta},$$

where π is the prior on $\boldsymbol{\beta}$. For example, with $\pi(\boldsymbol{\beta})$ is the independent normal $\text{N}(0, \sigma_k^2)$ on the k-th element of $\boldsymbol{\beta}$, let $h_\nu(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \Sigma^{-1}\boldsymbol{\beta}/2$, where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, we have

$$\frac{\partial h_\nu(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial \boldsymbol{\beta}} - \Sigma^{-1}\boldsymbol{\beta}$$

$$\frac{\partial^2 h_\nu(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = \frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} - \Sigma^{-1}.$$

Hence the Newton-Raphson step for solving the MLE of $\boldsymbol{\beta}$ given ν is given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \left(\boldsymbol{X}^\top W_2^{(t)} \boldsymbol{X} + \Sigma^{-1}/\nu\right)^{-1} \left(\boldsymbol{X}^\top W_1^{(t)} \boldsymbol{z}^{(t)} - \Sigma^{-1}\hat{\boldsymbol{\beta}}^{(t)}/\nu\right).$$

Under the alternative, suppose that $\pi(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta}_1)\pi(\gamma)$, where $\boldsymbol{\beta}_1$ are the coefficients for the covariates and γ for the group indicator. Instead of using independent normal prior on γ, the LIM g-prior (Li and Clyde, 2015) could be adopted. Using the independent normal prior for $\boldsymbol{\beta}_1$, we have

$$\begin{aligned} h_\nu(\boldsymbol{\beta}) &= l(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}) \\ &= l(\boldsymbol{\beta}) - \boldsymbol{\beta}_1^\top \Sigma^{-1}\boldsymbol{\beta}_1/2 - g^{-1}\mathcal{J}_\nu(\hat{\gamma})\gamma^2/2 \\ &= l(\boldsymbol{\beta}) - \boldsymbol{\beta}_1^\top \Sigma^{-1}\boldsymbol{\beta}_1/2 - g^{-1}\nu(\boldsymbol{X}^\top \hat{W}_2 \boldsymbol{X})_2 \gamma/2 \end{aligned}$$

where $(\boldsymbol{X}^\top \hat{W}_2 \boldsymbol{X})_2$ denote the block of the Hessian matrix corresponding to γ. There-

fore,

$$\frac{\partial h_\nu(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \nu \boldsymbol{X}^\top W_1 \boldsymbol{z} - \begin{pmatrix} \Sigma^{-1}\boldsymbol{\beta}_1 \\ g^{-1}\nu(\boldsymbol{X}^\top \hat{W}_2 \boldsymbol{X})_2 \gamma \end{pmatrix}$$

$$\frac{\partial^2 h_\nu(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\nu \boldsymbol{X}^\top W_2 \boldsymbol{X} - \begin{pmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & g^{-1}\nu(\boldsymbol{X}^\top \hat{W}_2 \boldsymbol{X})_2 \end{pmatrix}.$$

The resulting NR update is

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \left(\boldsymbol{X}^\top W_2^{(t)} \boldsymbol{X} + \begin{pmatrix} \Sigma^{-1}/\nu & \mathbf{0} \\ \mathbf{0} & g^{-1}(\boldsymbol{X}^\top W_2^{(t)} \boldsymbol{X})_2 \end{pmatrix} \right)^{-1} \times \left(\boldsymbol{X}^\top W_1^{(t)} \boldsymbol{z}^{(t)} - \begin{pmatrix} \hat{\Sigma}^{-1}\boldsymbol{\beta}_1^{(t)}/\nu \\ g^{-1}(\boldsymbol{X}^\top W_2^{(t)} \boldsymbol{X})_2 \hat{\gamma}^{(t)} \end{pmatrix} \right).$$

A.2 More on decision making

We first consider the original hypothesis that there is no cross-group difference. Taking a decision theoretic perspective, let $d(\boldsymbol{y}) \in \{0, 1\}$ be some decision rule, with $d(\boldsymbol{y}) = 1$ corresponding to the rejection of the global null that there are no cross-group differences in the OTU composition. When the loss function is

$$L(d(\boldsymbol{y}), c) = c \cdot \mathbb{1}_{[H_0 \text{ is true}]} d(\boldsymbol{y}) + (1 - c) \cdot \mathbb{1}_{[H_1 \text{ is true}]} (1 - d(\boldsymbol{y}))$$

for some $0 \leqslant c \leqslant 1$, one can show that the Bayes optimal decision rule is $d(\boldsymbol{y}) = \mathbb{1}_{[\text{PJAP} > c]}$. In particular, when $c = 0.5$, this gives the optimal decision under the simple 0-1 loss.

The decision on reporting the significant nodes is essentially a multiple testing problem. One way to address this problem is to use loss functions specified with the false positives and false negatives (Müller et al., 2006). For example, let $d_i(\boldsymbol{y}) \in \{0, 1\}$ be the decision rule on the i-th node; again, $d_i(\boldsymbol{y}) = 1$ corresponds to the rejection of the node-specific null. Let FD and FN denote the number of false positives and

false negatives. The posterior expectation of FD and FN are

$$\overline{\text{FD}} = \sum (1 - \text{PMAP}_i) \times d_i(\boldsymbol{y}),$$

$$\overline{\text{FN}} = \sum \text{PMAP}_i \times (1 - d_i(\boldsymbol{y})).$$

It can be shown that under the loss $L(d(\boldsymbol{y}), t) = t \times \text{FD} + \text{FN}$, the Bayes optimal decision rule, which minimizes the posterior expected loss $\overline{L}(d(\boldsymbol{y}), t) = t \times \overline{\text{FD}} + \overline{\text{FN}}$ has the form $d_i(\boldsymbol{y}) = \mathbb{1}_{[\text{PMAP}_i > c']}$ with the optimal threshold $c' = t/(t+1)$, $t \geqslant 0$ (Müller et al., 2004). In our application, we use $c' = 0.5$ that corresponds to $t = 1$ which is also recommended by Barbieri et al. (2004) from a Bayesian model choice perspective. Note that one can also consider loss functions that directly take into account the dependency among the hypotheses being tested. In our framework, such dependency is incorporated only through the probability model, not in the decision theoretic part.

A.3 Covariate selection

As we noted in Section 4.2.4, covariate selection is achievable in BGCR by putting a spike-and-slab prior on the regression coefficients (George and McCulloch, 1997). For example, let $r_l \overset{\text{ind}}{\sim} \text{Bernoulli}(q_l)$ where $q_l \in (0,1)$, $l = 2, \ldots, p+1$. For $A \in \mathcal{I}$, we can modify the prior on $\boldsymbol{\beta}(A)$ to be

$$\beta_l(A) \overset{\text{ind}}{\sim} (1 - r_l)\delta_0 + r_l \text{N}(0, \sigma_l^2(A)), \quad l = 2, \ldots, p+1, \tag{A.1}$$

where δ_0 is a point mass at zero, $\sigma_l^2(A)$'s are chosen for $\text{N}(0, \sigma_l^2(A))$ to cover all reasonable values of $\beta_l(A)$ while not supporting unreasonable values of $\beta_l(A)$.

Let $\boldsymbol{r} = (r_2, \ldots, r_{p+1}) \in \{0,1\}^p$. The independent Bernoulli priors on r_l induce the following prior on $\boldsymbol{r}$

$$\pi(\boldsymbol{r}) = \prod_{l=2}^{p+1} q_l^{r_l} (1 - q_l)^{1 - r_l}.$$

Conditioning on $\boldsymbol{r}$, the marginal likelihood of the data, $\phi_1(\Omega \mid \boldsymbol{r})$, is available as a byproduct of the BGCR inference algorithm (Section 4.2.4). When the number of covariates is not too large, this allows us to get the posterior of $\boldsymbol{r}$ by Bayes theorem:

$$\pi(\boldsymbol{r} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{r}) \phi_1(\Omega \mid \boldsymbol{r}).$$

We modify our simulation scenario IV in Section 3.3.3 to give a simple illustration of the covariate selection procedure. Consider the data simulated under the alternative, in which the counts of OTU '4481131' (ω_s) are increased by 175% in the second group. Instead of using "gender" as a confounder, we generate two covariates for each sample:

$$x_{ij2} \overset{\text{iid}}{\sim} \text{N}(0,1), \quad x_{ij3} \overset{\text{iid}}{\sim} \text{N}(0,1).$$

Suppose that the first covariate is relevant to the counts of a specific OTU while the second covariate has nothing to do with the OTU counts. Specifically, we increase the counts of OTU '4352657' (ω_c) in the j-th sample in group i by $(x_{ij2} \times 175\%)$ (when this value is less than -1, we set the count to zero). We note that due to the large variation in OTU counts, the signal injected on ω_c is quite weak.

Consider a specific round of simulation. We let $r_2 = r_3 = 0.5$, $\sigma_l^2(A) = 10$ for $l = 2, 3$ and fit BGCR with the prior in (A.1). A.1 summaries the posterior probabilities of the four possible models. In comparison, each model has equal prior probabilities. Therefore, the important variable is correctly identified.

Table A.1: Posterior probabilities of different models (no confounding).

Covariate in the model	None	2	3	2 and 3
Posterior probability	0.223	0.320	0.186	0.270

Although a covariate selection procedure can be incorporated in BGCR, one must proceed with caution since this can substantially affect or even invalidate the

meaning of the testing result on the two-group difference. To see this intuitively, consider the following simplistic but representative scenario. Suppose there is a (close-to) perfect confounding covariate which explains virtually all the difference across the two groups. Once this covariate is included in the model then there is no remaining cross-group difference and the two-group comparison will not favor the alternative. However, including the covariate into the model may not improve the fit to the observed data in any substantive manner as its effect is largely overlapping with that of the intercept (i.e., the group label). Consequently, statistical model selection strategies, both Bayesian or frequentist, would very likely to exclude this covariate from the model. This would lead to a significant testing result on the two-group differences. As a simple illustration, in the previous example, suppose instead we have

$$x_{1j2} \overset{\text{iid}}{\sim} \mathrm{N}(0,1), \quad x_{2j2} \overset{\text{iid}}{\sim} \mathrm{N}(2,1).$$

In this case, the first covariate is a confounding variable. A.2 summaries the posterior probabilities of the four possible models. Due to the strong confounding effect, the first covariate is excluded from the model, which would lead to false positives in the testing scenario.

Table A.2: Posterior probabilities of different models (with confounding).

Covariate in the model	None	2	3	2 and 3
Posterior probability	0.599	0.001	0.400	≈ 0

A.4 Additional figures

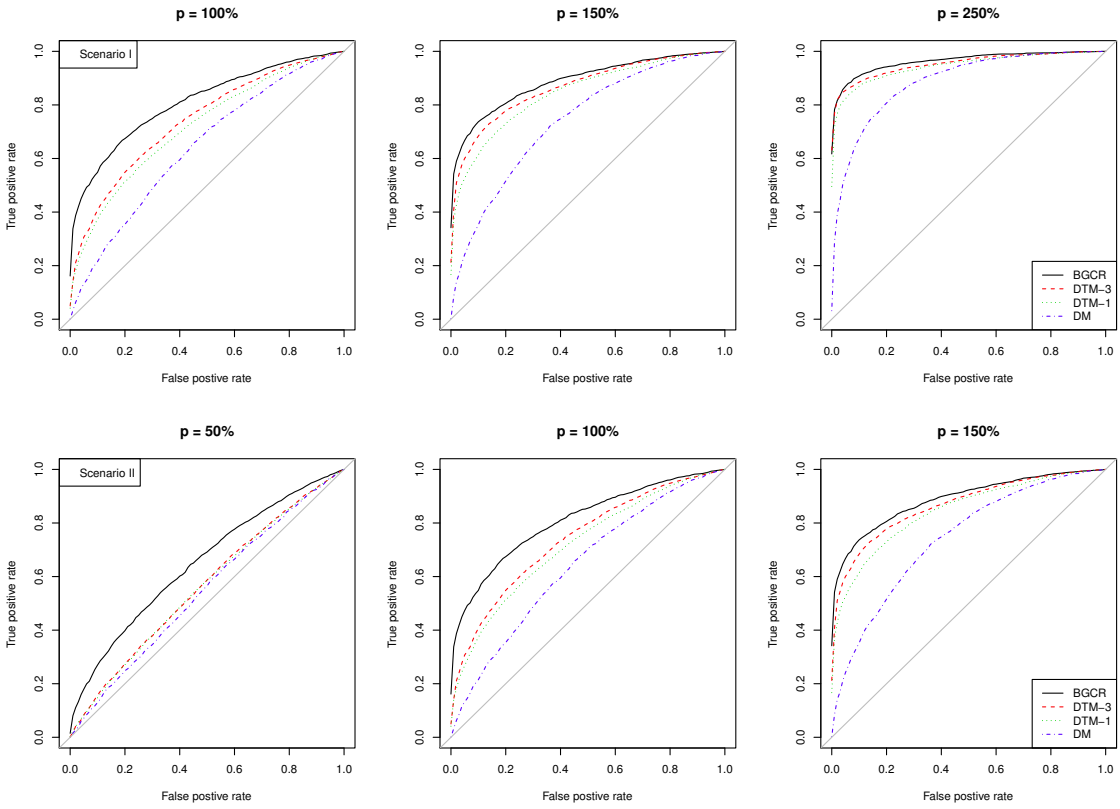

FIGURE A.1: ROC curves for Scenario I and II with $M = 50$. The columns are indicated by the percent of count increased in the second group (p).

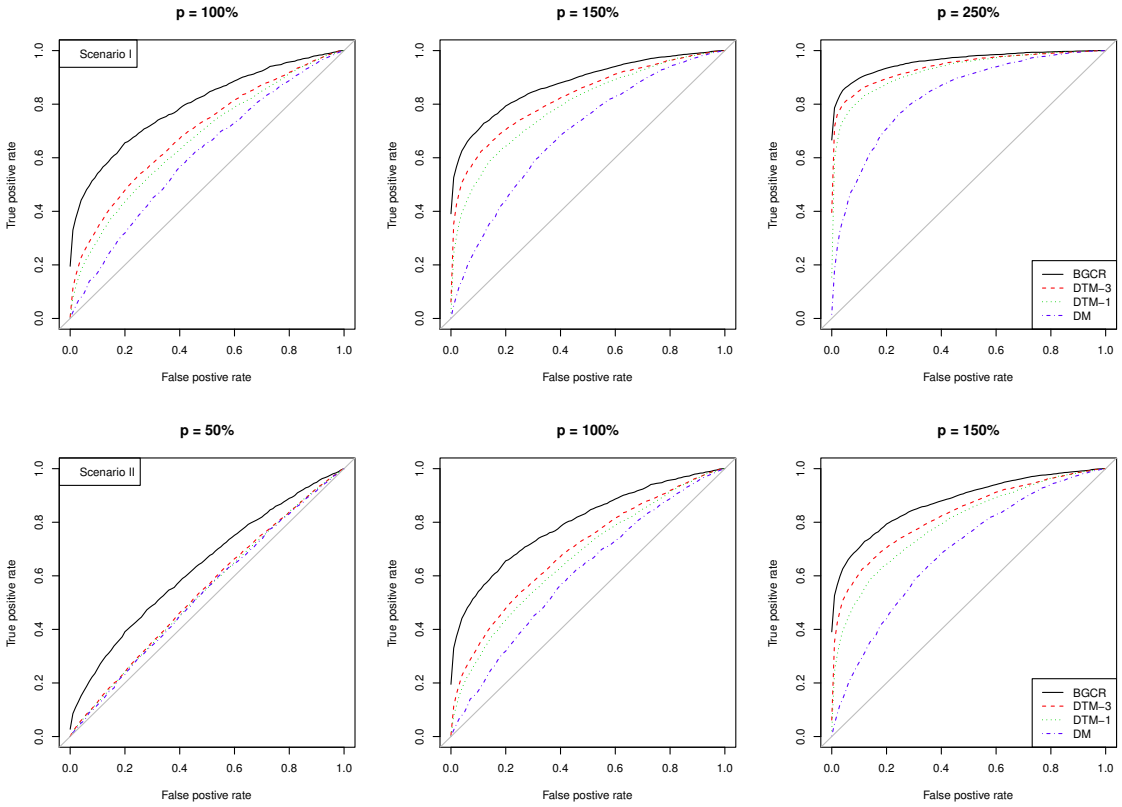

FIGURE A.2: ROC curves for Scenario I and II with $M = 75$. The columns are indicated by the percent of count increased in the second group (p).

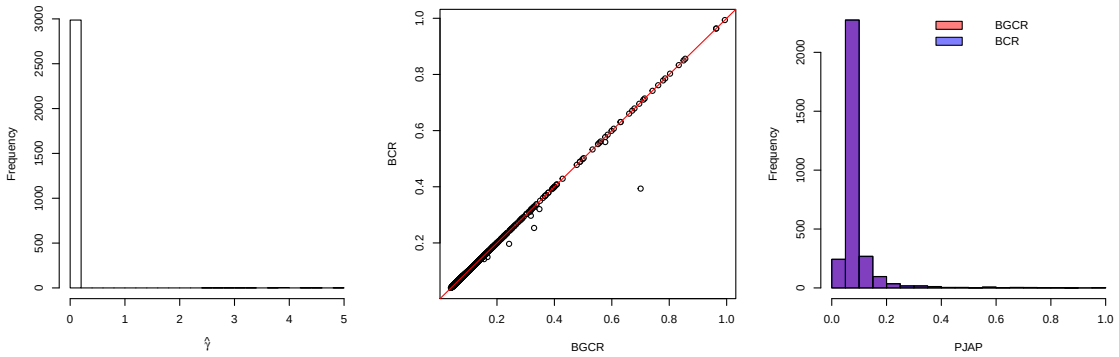

FIGURE A.3: BGCR vs BCR under the null in scenario 3. Left: Histogram of the estimated γ in BGCR; Middle: PJAPs of BGCR vs BCR; Right: Histograms of the PJAPs of BGCR and BCR.

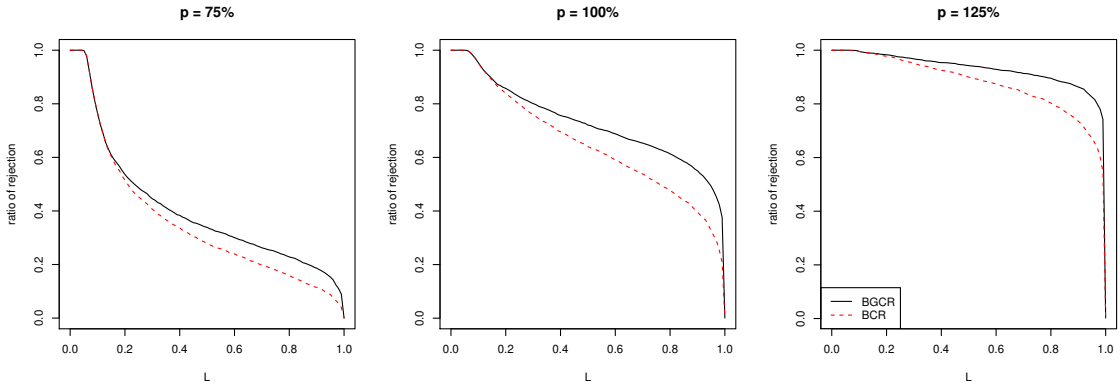

FIGURE A.4: Ratio of rejection under the alternatives in Scenario III. The columns are indicated by the percent of count increased in the second group (p).

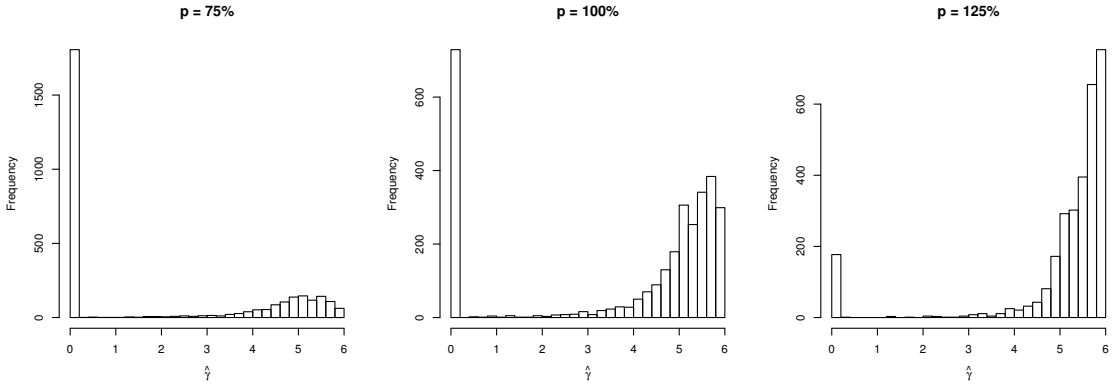

FIGURE A.5: Estimated γ under the alternatives in scenario 3.

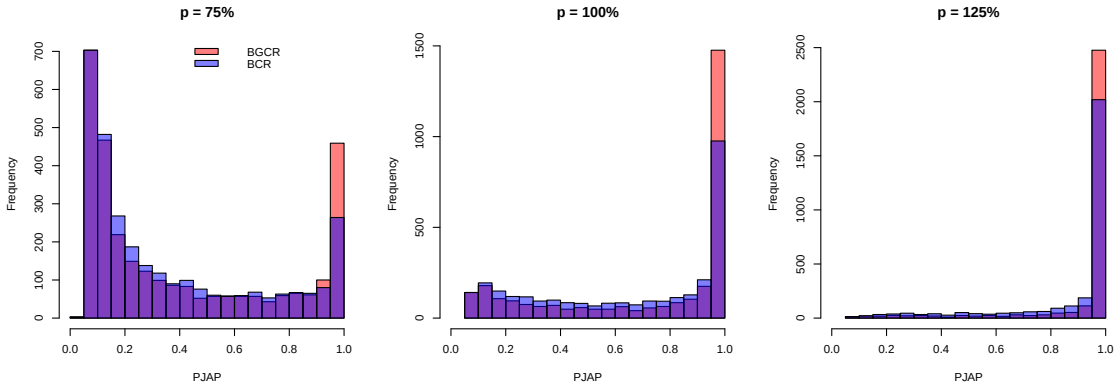

FIGURE A.6: Histograms of the PJAPs under the alternatives in scenario 3.

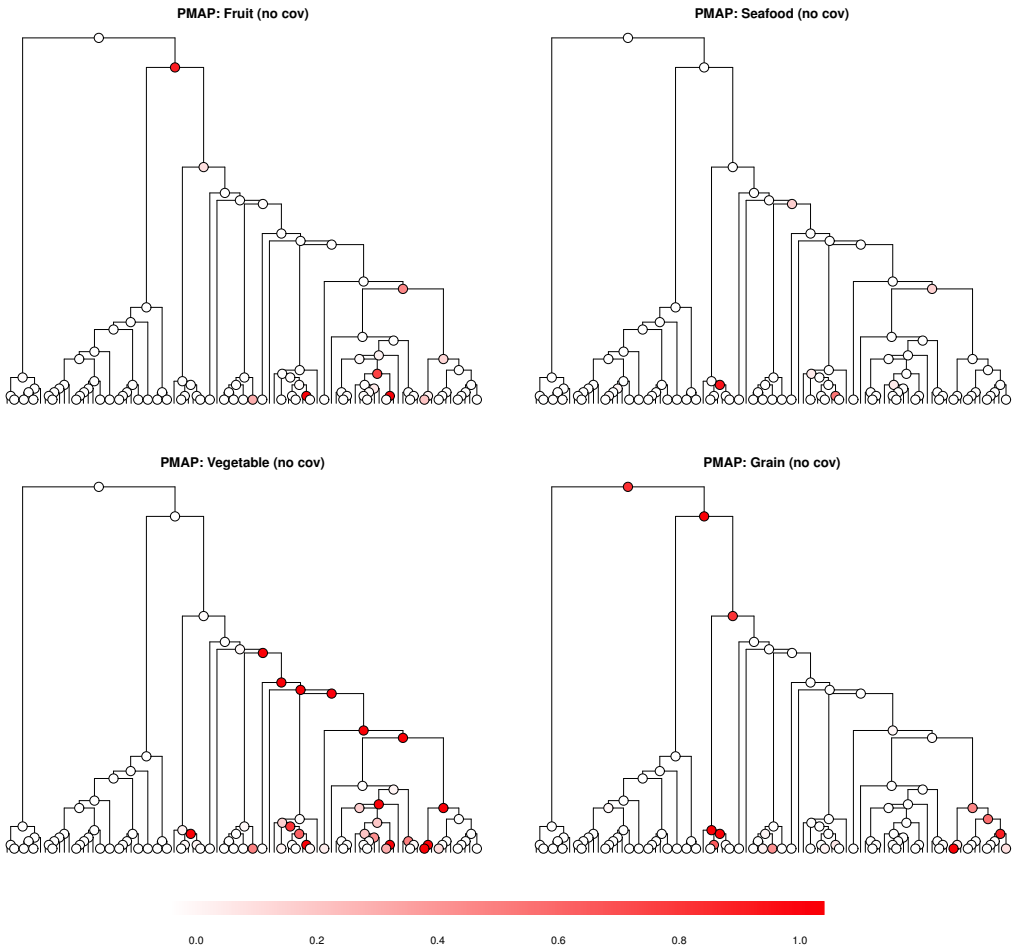

FIGURE A.7: PMAPs for the four comparisons that reject the global null. The nodes are colored by PMAPs reported by BGCR with no covariate adjusted.

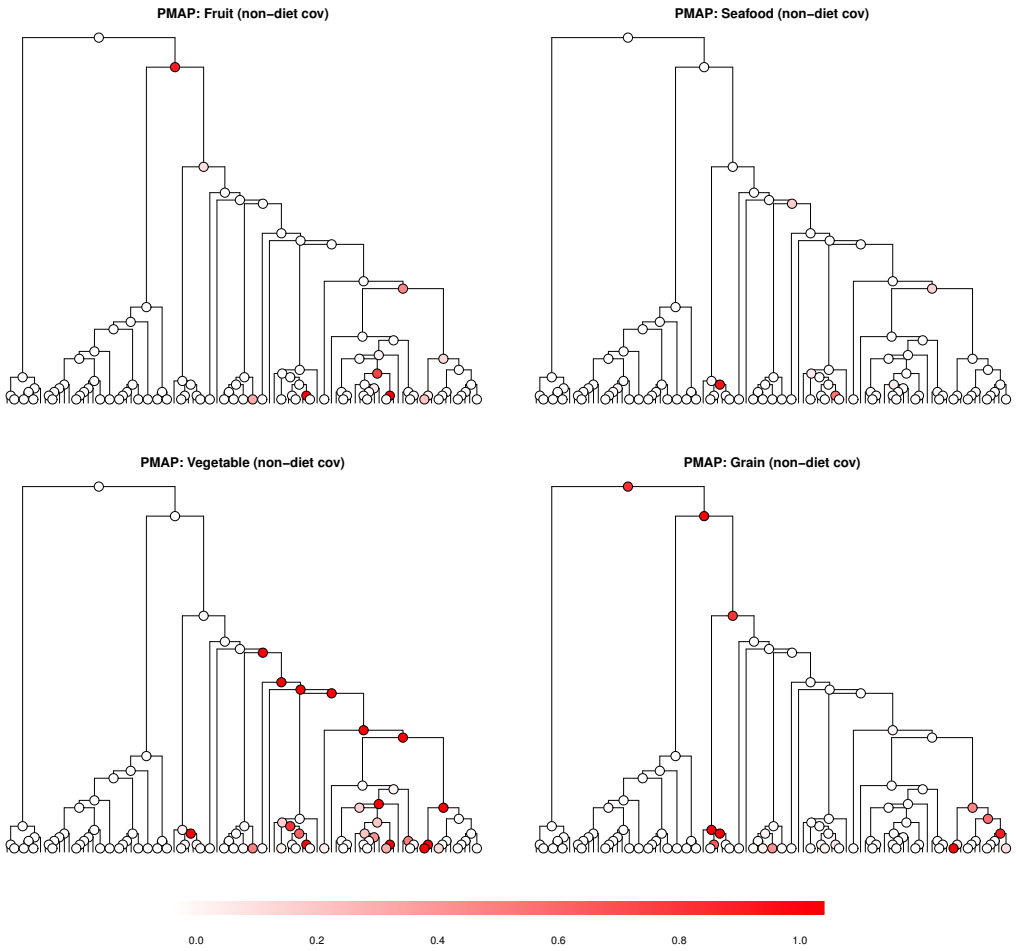

FIGURE A.8: PMAPs for the four comparisons that reject the global null. The nodes are colored by PMAPs reported by BGCR with only non-dietary covariates adjusted.

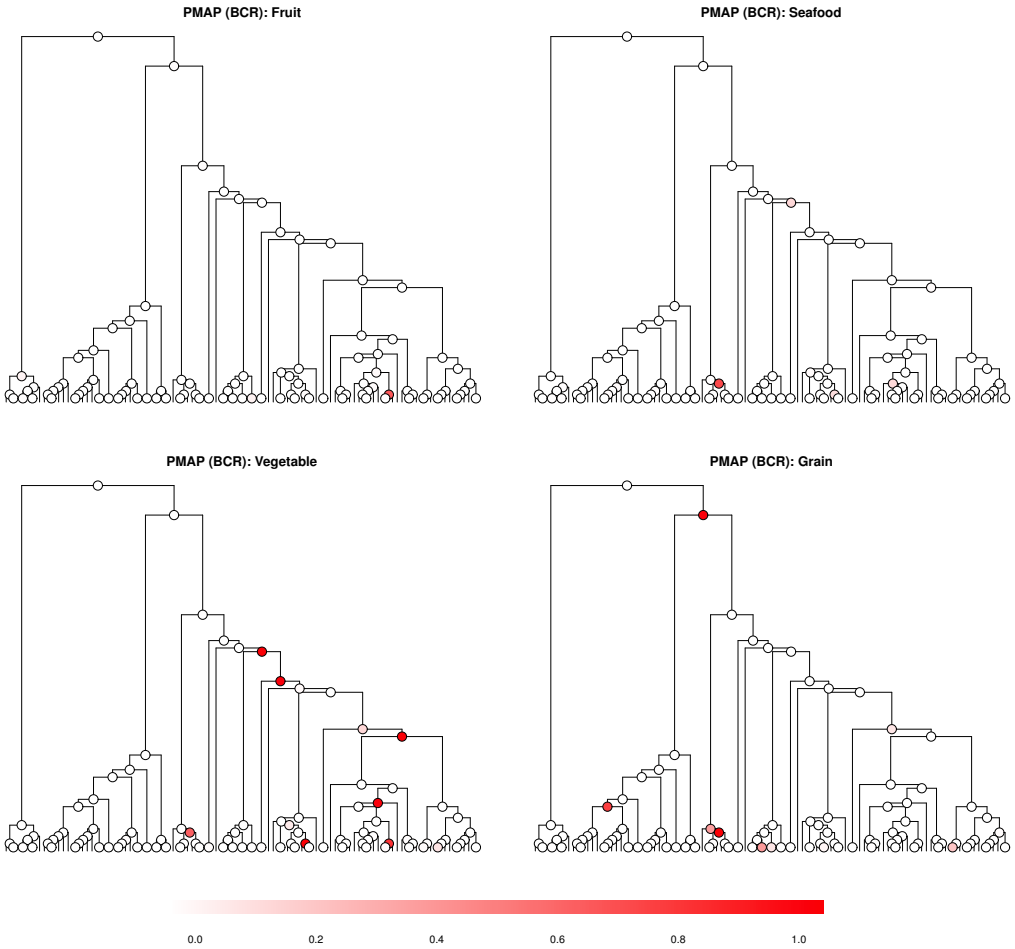

FIGURE A.9: PMAPs for the four comparisons that reject the global null. The nodes are colored by PMAPs reported by BCR with both non-dietary covariates and dietary covariates adjusted.

Appendix B

Supplementary Material for Chapter 4

B.1 Computational strategies

B.1.1 Derivation of the Gibbs sampler for DTMM.

The key to the derivation of the Gibbs sampler is to marginalize out the sample and cluster-specific parameters such that the number of parameters need to be sampled is greatly reduce.

Recall that for $\boldsymbol{Y}_c^I = \{\boldsymbol{y}_i : c_i = c, i \in I\}$, the marginal likelihood of $\boldsymbol{Y}_c^I$ at node A given $\gamma(A) = 1$ or $\gamma(A) = 0$ after marginalizing out the sample-specific parameter and the cluster-specific parameter can be written as

$$\begin{aligned}\mathcal{L}_1^A(\boldsymbol{Y}_c^I) &:= \iint \mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)) d\Pi(\boldsymbol{\psi}_c^*(A) \mid \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)) \\ &= \iint \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta(A)\tau(A) + y_i(A_l), (1-\theta(A))\tau(A) + y_i(A_r))}{B(\theta(A)\tau(A), (1-\theta(A))\tau(A))} \\ &\quad \times \frac{\theta(A)^{\theta_0(A)\nu_0(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu_0(A)-1}}{B(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A))} d\theta(A) dF^A(\tau),\end{aligned} \tag{B.1}$$

$$\mathcal{L}_0^A(\boldsymbol{Y}_c^I \mid \tilde{\boldsymbol{\psi}}(A)) := \iint \mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A)=0, \tilde{\boldsymbol{\psi}}(A)) d\Pi(\boldsymbol{\psi}_c^*(A) \mid \gamma(A)=0, \tilde{\boldsymbol{\psi}}(A))$$

$$= \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\tilde{\theta}(A)\tilde{\tau}(A) + y_i(A_l), (1-\tilde{\theta}(A))\tilde{\tau}(A) + y_i(A_r))}{B(\tilde{\theta}(A)\tilde{\tau}(A), (1-\tilde{\theta}(A))\tilde{\tau}(A))}. \tag{B.2}$$

(i). **The full conditional of $\boldsymbol{\gamma}$.** For each $A \in \mathcal{I}$, let $\boldsymbol{Y}(A) = \{\boldsymbol{y}_i(A) : i \in [n]\}$ and $\boldsymbol{Y}_c(A) = \{\boldsymbol{y}_i(A) : c_i = c, i \in [n]\}$. We have

$$\begin{aligned}
& M_{10}(A \mid \boldsymbol{\gamma}^{-A}, \boldsymbol{c}, \beta, \lambda) \\
&= \frac{\mathcal{L}(\boldsymbol{Y} \mid \gamma(A)=1, \boldsymbol{\gamma}^{-A}, \boldsymbol{c}, \beta, \lambda)}{\mathcal{L}(\boldsymbol{Y} \mid \gamma(A)=0, \boldsymbol{\gamma}^{-A}, \boldsymbol{c}, \beta, \lambda)} \\
&= \frac{\mathcal{L}(\boldsymbol{Y}(A_l) \mid \boldsymbol{Y}(A), \gamma(A)=1, \boldsymbol{c}, \beta, \lambda)}{\mathcal{L}(\boldsymbol{Y}(A_l) \mid \boldsymbol{Y}(A), \gamma(A)=0, \boldsymbol{c}, \beta, \lambda)} \\
&= \frac{\prod_{c \in \boldsymbol{c}^*} \int \mathcal{L}(\boldsymbol{Y}_c(A_l) \mid \boldsymbol{Y}_c(A), \boldsymbol{\psi}_c^*(A), \gamma(A)=1) d\Pi(\boldsymbol{\psi}_c^*(A) \mid \gamma(A)=1)}{\int \mathcal{L}(\boldsymbol{Y}(A_l) \mid \boldsymbol{Y}(A), \tilde{\boldsymbol{\psi}}(A), \gamma(A)=0) d\Pi(\tilde{\boldsymbol{\psi}}(A))} \\
&= \frac{\prod_{c \in \boldsymbol{c}^*} \mathcal{L}_1^A(\boldsymbol{Y}_c)}{\int \mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A)) d\Pi(\tilde{\boldsymbol{\psi}}(A))}.
\end{aligned} \tag{B.3}$$

(ii). **The full conditional of $\boldsymbol{c}$.** For $i \in [n]$, let $\boldsymbol{Y}^{-i}$ denote the set of samples with sample i excluded and let $\boldsymbol{c}_{-i}^* = \{\boldsymbol{c}_{-i}\}$ be the set of distinct values of $\boldsymbol{c}_{-i}$.

For $c \in \boldsymbol{c}^*$, we have

$$
\begin{aligned}
&\Pr(c_i = c \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) \\
\propto &\Pr(c_i = c \mid \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, c_i = c, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) \\
\propto &n_{-i,c} \times \int \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, c_i = c, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda, \tilde{\boldsymbol{\psi}}) d\Pi(\tilde{\boldsymbol{\psi}}) \\
\propto &n_{-i,c} \times \int \left[\int \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, \boldsymbol{\psi}_c^*, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda, \tilde{\boldsymbol{\psi}}) d\Pi(\boldsymbol{\psi}_c^* \mid \tilde{\boldsymbol{\psi}}, \boldsymbol{\gamma}) \right] d\Pi(\tilde{\boldsymbol{\psi}}) \\
\propto &n_{-i,c} \times \prod_{\{A \in \mathcal{I}: \gamma(A) = 1\}} \frac{\mathcal{L}_1^A(\boldsymbol{y}_i, \boldsymbol{Y}_c^{-i})}{\mathcal{L}_1^A(\boldsymbol{Y}_c^{-i})} \times \left[\int \prod_{\{A \in \mathcal{I}: \gamma(A) = 0\}} \frac{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y}^{-i} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}) \right] \\
\propto &n_{-i,c} \times \frac{\mathcal{L}_1(\boldsymbol{y}_i, \boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})}{\mathcal{L}_1(\boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})} \times \left[\int \prod_{\{A \in \mathcal{I}: \gamma(A) = 0\}} \frac{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y}^{-i} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}) \right],
\end{aligned}
\tag{B.4}
$$

where we use $\mathcal{L}(\boldsymbol{y}_i \mid -)$ to denote the conditional likelihood of $\boldsymbol{y}_i$ given certain parameters or other samples. Similarly, we have

$$
\begin{aligned}
&\Pr(c_i \neq c_j \text{ for all } j \neq i \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) \\
\propto &\Pr(c_i \notin \boldsymbol{c}_{-i}^* \mid \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, c_i \notin \boldsymbol{c}_{-i}^*, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda) \\
\propto &\beta \times \int \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, c_i \notin \boldsymbol{c}_{-i}^*, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda, \tilde{\boldsymbol{\psi}}) d\Pi(\tilde{\boldsymbol{\psi}}) \\
\propto &\beta \times \int \left[\int \mathcal{L}(\boldsymbol{y}_i \mid \boldsymbol{Y}^{-i}, \boldsymbol{\psi}_{c_i}^*, \boldsymbol{c}_{-i}, \boldsymbol{\gamma}, \beta, \lambda, \tilde{\boldsymbol{\psi}}) d\Pi(\boldsymbol{\psi}_{c_i}^* \mid \tilde{\boldsymbol{\psi}}, \boldsymbol{\gamma}) \right] d\Pi(\tilde{\boldsymbol{\psi}}) \\
\propto &\beta \times \prod_{\{A \in \mathcal{I}: \gamma(A) = 1\}} \mathcal{L}_1^A(\boldsymbol{y}_i) \times \left[\int \prod_{\{A \in \mathcal{I}: \gamma(A) = 0\}} \frac{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y}^{-i} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}) \right] \\
\propto &\beta \times \mathcal{L}_1(\boldsymbol{y}_i \mid \boldsymbol{\gamma}) \times \left[\int \prod_{\{A \in \mathcal{I}: \gamma(A) = 0\}} \frac{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y}^{-i} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}) \right].
\end{aligned}
\tag{B.5}
$$

Putting together (B.4) and (B.5), we have

$$\begin{aligned}\Pr(c_i = c \text{ for some } c \in c_{-i} \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) &\propto n_{-i,c} \times \frac{\mathcal{L}_1(\boldsymbol{y}_i, \boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})}{\mathcal{L}_1(\boldsymbol{Y}_c^{-i} \mid \boldsymbol{\gamma})} \\ \Pr(c_i \neq c_j \text{ for all } j \neq i \mid \boldsymbol{c}_{-i}, \boldsymbol{Y}, \boldsymbol{\gamma}, \beta, \lambda) &\propto \beta \times \mathcal{L}_1(\boldsymbol{y}_i \mid \boldsymbol{\gamma}).\end{aligned} \tag{B.6}$$

Note that the integral term in (B.4) and (B.5) represents the marginal likelihood of $\boldsymbol{y}_i$ at the nodes with $\gamma(A) = 0$ conditioning on all other samples. This likelihood does not depend on $\boldsymbol{c}_{-i}$ since the cluster-specific parameters $\boldsymbol{\psi}_c^*$ share the same value at these nodes regardless of c.

B.1.2 Numerical approximations to the marginal likelihoods

To fully specify the Gibbs sampler, we need to numerically approximate the following marginal likelihoods at $A \in \mathcal{I}$:

1. The marginal likelihoods of a set of samples from the same cluster given $\gamma(A) = 1$: $\mathcal{L}_1^A(\boldsymbol{Y}_c^I)$;

2. The marginal likelihoods of the samples given $\gamma(A) = 0$:

$$\int \mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A)) d\Pi(\tilde{\boldsymbol{\psi}}(A)). \tag{B.7}$$

Note that (B.7) is a one-dimensional integral that can be easily approximated by quadrature. For $\mathcal{L}_1^A(\boldsymbol{Y}_c^I)$, let

$$\begin{aligned} g(\theta, \tau) &= \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta\tau + y_i(A_l), (1-\theta)\tau + y_i(A_r))}{B(\theta\tau, (1-\theta)\tau)} \\ h(\theta) &= \frac{\theta^{\theta_0(A)\nu_0(A)-1}(1-\theta)^{(1-\theta_0(A))\nu_0(A)-1}}{B(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A))}. \end{aligned} \tag{B.8}$$

Then we have

$$\mathcal{L}_1^A(\boldsymbol{Y}_c^I) = \int \left[\int g(\theta, \tau) dF^A(\tau) \right] h(\theta) d\theta. \tag{B.9}$$

Both integrals in (B.9) are one-dimensional. In our software, we approximate both integrals with quadratures.

B.1.3 Sample classification for microbiome data

In this section, we provide details on performing sample classification for microbiome data under the DTMM framework. Let $\tilde{\boldsymbol{y}}$ be a new microbiome sample from (2.33) in section 2.5. Similar to the computations in (B.4), we have

$$
\begin{aligned}
&\Pr(\tilde{c}=k \mid \tilde{\boldsymbol{y}}, \boldsymbol{Y}) \\
&\propto \pi_k \mathcal{L}(\tilde{\boldsymbol{y}} \mid \tilde{c}=k, \boldsymbol{Y}) \\
&\propto \pi_k \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \iint \mathcal{L}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\psi}_k^*, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}, \boldsymbol{Y}_k) d\Pi(\boldsymbol{\psi}_k^* \mid \boldsymbol{Y}_k, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}) d\Pi(\tilde{\boldsymbol{\psi}} \mid \boldsymbol{Y}) \\
&\propto \pi_k \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \int \left[\int \mathcal{L}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\psi}_k^*, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}, \boldsymbol{Y}_k) d\Pi(\boldsymbol{\psi}_k^* \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}) \right] d\Pi(\tilde{\boldsymbol{\psi}}) \\
&\propto \pi_k \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \prod_{\{A \in \mathcal{I}: \gamma(A)=1\}} \frac{\mathcal{L}_1^A(\tilde{\boldsymbol{y}}, \boldsymbol{Y}_k)}{\mathcal{L}_1^A(\boldsymbol{Y}_k)} \times \left[\int \prod_{\{A \in \mathcal{I}: \gamma(A)=0\}} \frac{\mathcal{L}_0^A(\tilde{\boldsymbol{y}}, \boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}) \right] \\
&\propto \pi_k \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \prod_{\{A \in \mathcal{I}: \gamma(A)=1\}} \frac{\mathcal{L}_1^A(\tilde{\boldsymbol{y}}, \boldsymbol{Y}_k)}{\mathcal{L}_1^A(\boldsymbol{Y}_k)} \times \prod_{\{A \in \mathcal{I}: \gamma(A)=0\}} \int \frac{\mathcal{L}_0^A(\tilde{\boldsymbol{y}}, \boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))}{\mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A))} d\Pi(\tilde{\boldsymbol{\psi}}(A)),
\end{aligned}
\tag{B.10}
$$

where

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\gamma}) \mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{\gamma}) \\
&\qquad \propto \pi(\boldsymbol{\gamma}) \prod_{\{A \in \mathcal{I}: \gamma(A)=1\}} \prod_{1 \leqslant k \leqslant K} \mathcal{L}_1^A(\boldsymbol{Y}_k) \times \prod_{\{A \in \mathcal{I}: \gamma(A)=0\}} \int \mathcal{L}_0^A(\boldsymbol{Y} \mid \tilde{\boldsymbol{\psi}}(A)) d\Pi(\tilde{\boldsymbol{\psi}}(A)).
\end{aligned}
\tag{B.11}
$$

Note that (B.10) and (B.11) can be evaluated numerically.

B.2 Additional materials for the numerical examples

B.2.1 Logistic normal approximations to the DT distribution

Lemma 3. *For $d \geqslant 2$ and $d \in \mathbb{N}^+$, let Q be a distribution on $\mathbb{S}^{d-1}$. Let $\mathbb{L}^{d-1}$ be the set of logistic-normal distributions on $\mathbb{S}^{d-1}$. We have*

$$P = \operatorname*{arg\,min}_{h \in \mathbb{L}^d} D_{KL}(Q \parallel h),$$

where

$$P \overset{d}{=} \text{Logit-Norm}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \quad \tilde{\boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{x}\sim Q}\left[\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)\right], \quad \tilde{\boldsymbol{\Sigma}} = \mathbb{V}_{\boldsymbol{x}\sim Q}\left[\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)\right].$$

Proof. For $h \sim \text{Logit-Norm}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \in \mathbb{L}^d$, the pdf of h can be written as

$$\begin{aligned} p_h(\boldsymbol{x}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \left(\prod_{j=1}^{d} x_j\right)^{-1} e^{-\frac{1}{2}\left\{\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)-\boldsymbol{\mu}\right\}^\top \boldsymbol{\Sigma}^{-1}\left\{\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)-\boldsymbol{\mu}\right\}} \\ &= f(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \text{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}) \end{pmatrix} \\ \boldsymbol{T}(\boldsymbol{x}) &= \begin{pmatrix} \log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right) \\ \text{vec}\left[\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)^\top\right] \end{pmatrix} \\ A(\boldsymbol{\eta}) &= \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log\boldsymbol{\Sigma} \end{aligned}$$

Thus h is an exponential family distribution with natural parameter $\boldsymbol{\eta}$. Note that

$$\begin{aligned} D_{\text{KL}}(Q \parallel h) &= \mathbb{E}_Q\left[\log\left(\frac{Q}{h}\right)\right] \\ &= \mathbb{E}_Q\left[\log Q\right] + \mathbb{E}_Q\left[A(\boldsymbol{\eta})\right] - \mathbb{E}_Q\left[\boldsymbol{\eta}^\top\boldsymbol{T}(\boldsymbol{x})\right] - \mathbb{E}_Q\left[\log f(\boldsymbol{x})\right] \end{aligned}$$

By the property of the exponential family, we have

$$\nabla_{\boldsymbol{\eta}} D_{\mathrm{KL}}(Q \parallel h) = \mathbb{E}_h[\boldsymbol{T}(\boldsymbol{x})] - \mathbb{E}_Q\left[\boldsymbol{T}(\boldsymbol{x})\right].$$

Let h^* satisfy $\nabla_{\boldsymbol{\eta}} D_{\mathrm{KL}}(Q \parallel h) = 0$, if follows that $\mathbb{E}_{h^*}[\boldsymbol{T}(\boldsymbol{x})] = \mathbb{E}_Q\left[\boldsymbol{T}(\boldsymbol{x})\right]$. Consider the second derivative of $D_{\mathrm{KL}}(Q \parallel h)$ at h^*

$$\nabla\nabla_{\eta} D_{\mathrm{KL}}(Q \parallel h) = \nabla\nabla_{\eta}\mathbb{E}_Q\left[A(\boldsymbol{\eta})\right] = \mathbb{V}_{h^*}[\boldsymbol{T}(\boldsymbol{x})],$$

which is positive semi-definite. Therefore, h^* minimizes $D_{\mathrm{KL}}(Q \parallel h)$. □

Therefore, for $Q \overset{d}{=} \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\tau})$, the best approximation to Q in the logistic-normal family is $P \overset{d}{=} \text{Logit-Norm}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ such that

$$\tilde{\boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{x}\sim Q}\left[\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)\right], \quad \tilde{\boldsymbol{\Sigma}} = \mathbb{V}_{\boldsymbol{x}\sim Q}\left[\log\left(\frac{\boldsymbol{x}_{-d}}{x_d}\right)\right].$$

For the j-the OTU ω_j, let A_{jp} be its parent in the phylogenetic tree. Let $\alpha_j = \theta(A_{jp})\tau(A_{jp})$ if ω_j is the left child of A_{jp} and $\alpha_j = (1-\theta(A_{jp}))\tau(A_{jp})$ otherwise. Similarly, for $A \in \mathcal{T}$, let $\beta_A = \theta(A_p)\tau(A_p) - \theta(A)$ if A is the left child of A_p and $\beta_A = (1-\theta(A_p))\tau(A_p) - \theta(A)$ otherwise. The density function of Q can be written as

$$p_Q(\boldsymbol{x}) = \left[\prod_{A\in\mathcal{T}} B(\theta(A)\tau(A), (1-\theta(A))\tau(A))\right]^{-1} \prod_{j=1}^{d} x_j^{\alpha_j - 1} \prod_{A\in\mathcal{T}} x_A^{\beta_A},$$

where $x_A = \sum_{j\in A} x_j$. Therefore, the Dirichlet-tree distribution is a member of the exponential family with natural parameter $\boldsymbol{\eta} = \{\{\alpha_j : 1 \leqslant j \leqslant d\}, \{\beta_A : A \in \mathcal{T}\}\}$. Let

$$A(\boldsymbol{\eta}) = \sum_{A\in\mathcal{T}} \log\left[B(\theta(A)\tau(A), (1-\theta(A))\tau(A))\right].$$

Again by the property of the exponential family, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ can be computed with the digamma and the trigamma functions.

B.2.2 Additional figures for the numerical examples

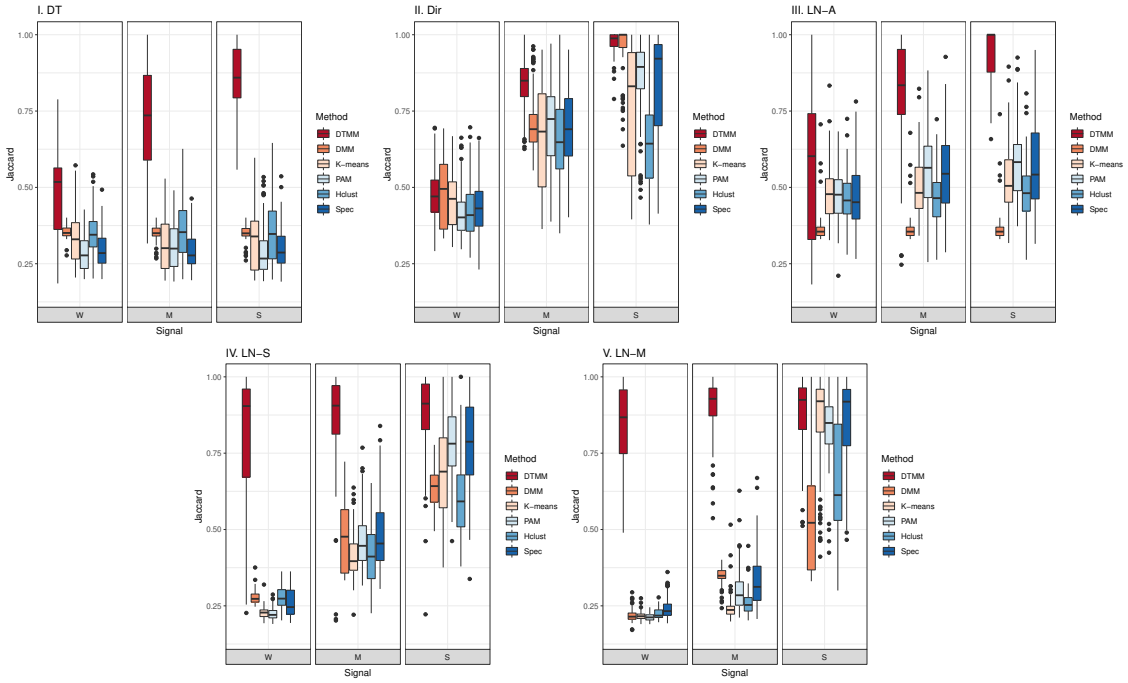

FIGURE B.1: Boxplots of the Jaccard index ($n = 90$).

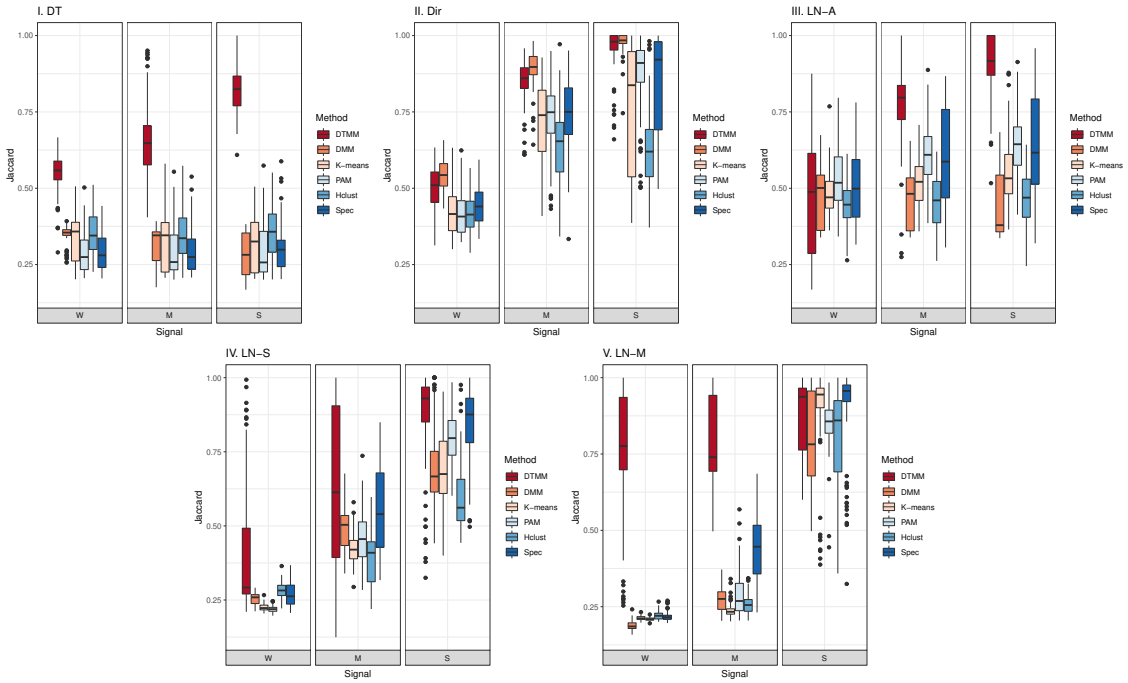

FIGURE B.2: Boxplots of the Jaccard index ($n = 180$).

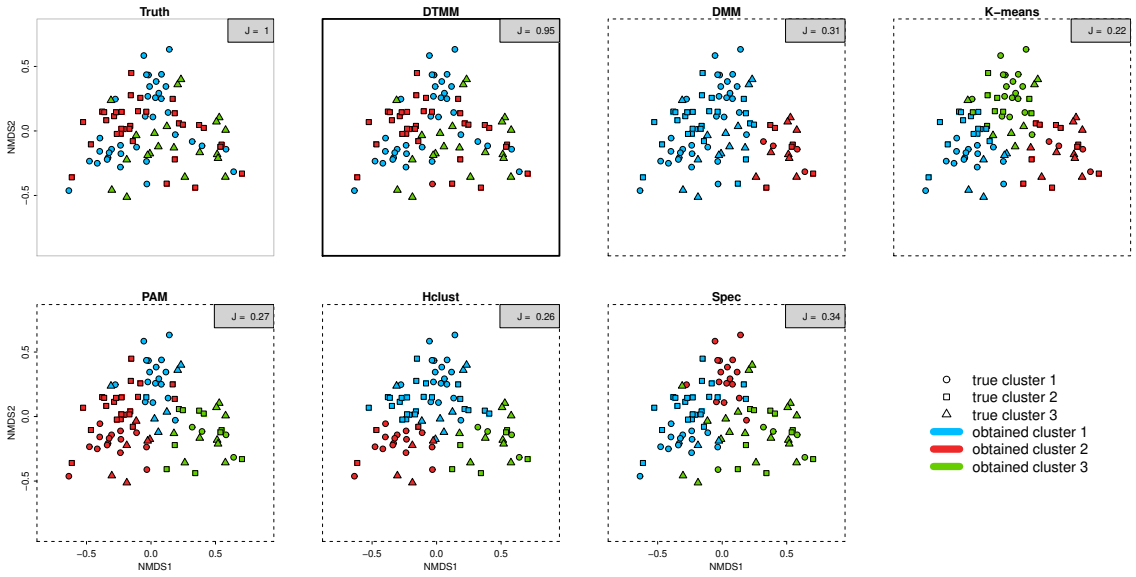

FIGURE B.3: 2D NMDS plot of samples in a simulation round in scenario V ($n = 90$, medium noise level). In each sub-plot, the true clustering is indicated by the shape of the points while the clustering obtained is indicated by the color.

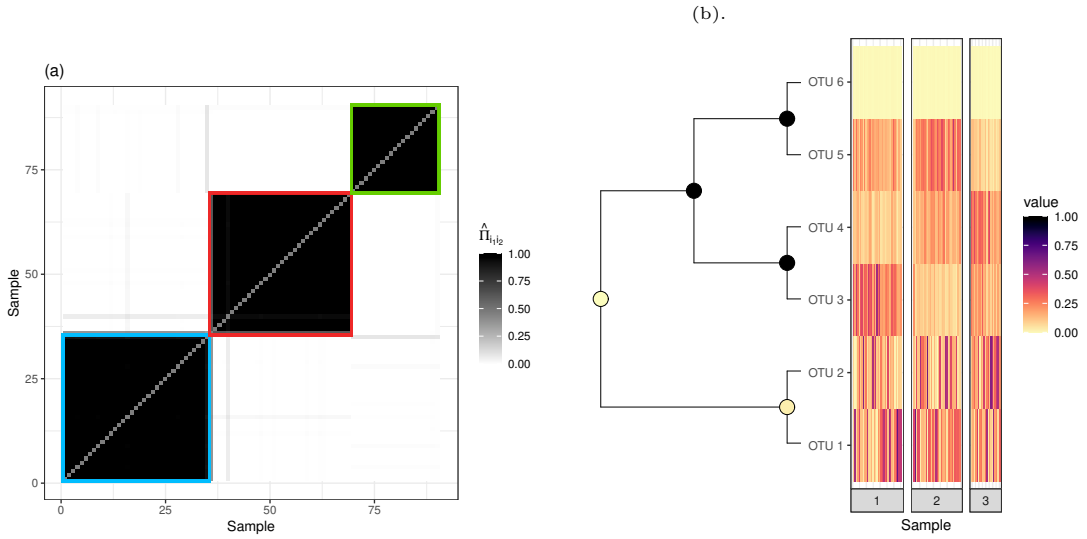

FIGURE B.4: Illustrations for an example from simulation scenario V. (a): Probability of two samples being clustered together by DTMM based on 1000 post-burnin MCMC samples. The samples are ordered by their cluster labels from DTMM. The clusters identified by DTMM are highlighted by squares colored as in 4.10. (b): An illustration of the node selection property of DTMM. The nodes are colored by their estimated posterior node selection probabilities. The heatmap plots the relative abundance of the samples grouped by their cluster labels from DTMM.

FIGURE B.5: The heatmap of the microbiome samples of patient D (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DMM. The legends are defined the same way as in 4.12.

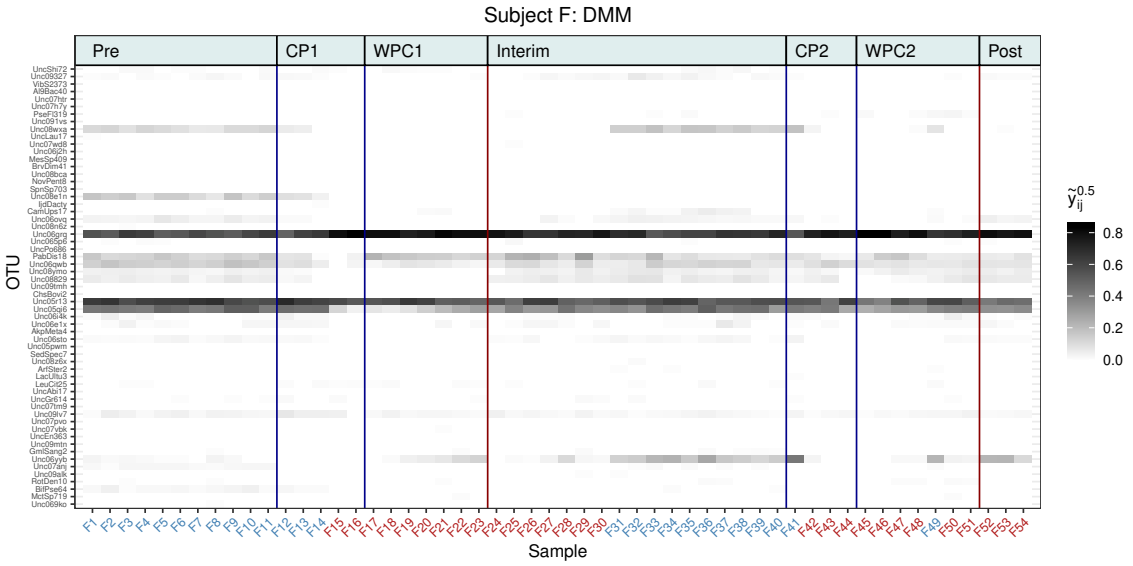

FIGURE B.6: The heatmap of the microbiome samples of patient F (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DMM. The legends are defined the same way as in 4.12.

B.3 Additional materials for the applications

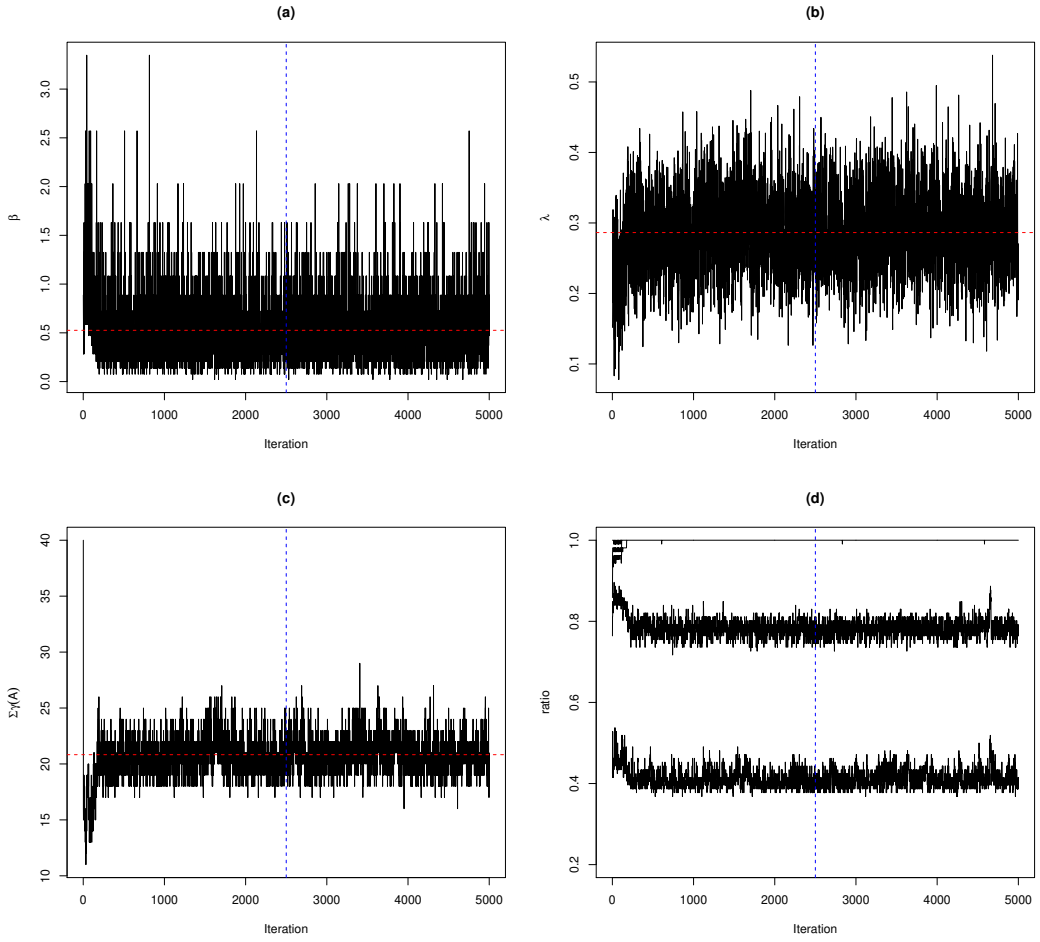

FIGURE B.7: Traceplots of some summary statistics for Section 4.2.

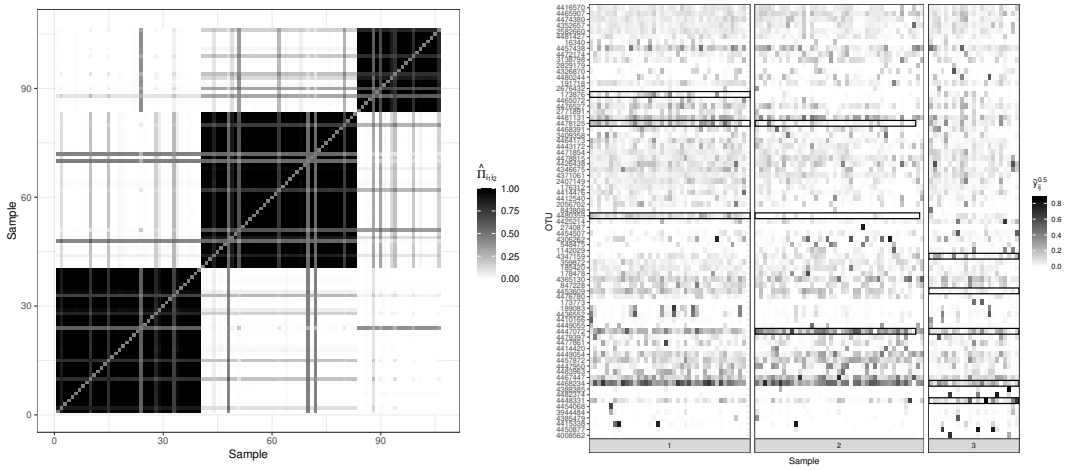

FIGURE B.8: Left: Estimated pairwise co-clustering probabilities. Right: Heatmap of the samples (after the square-root transform) grouped by their labels in $\boldsymbol{C}_{LS}$. The black boxes illustrate the characteristic OTUs of each cluster.

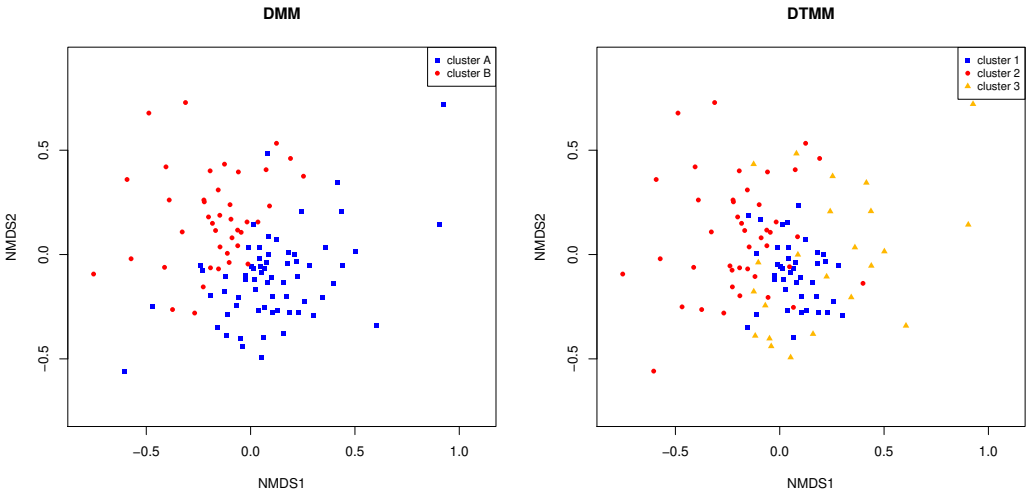

FIGURE B.9: Two-dimensional NMDS plots for the AG diabetes dataset.

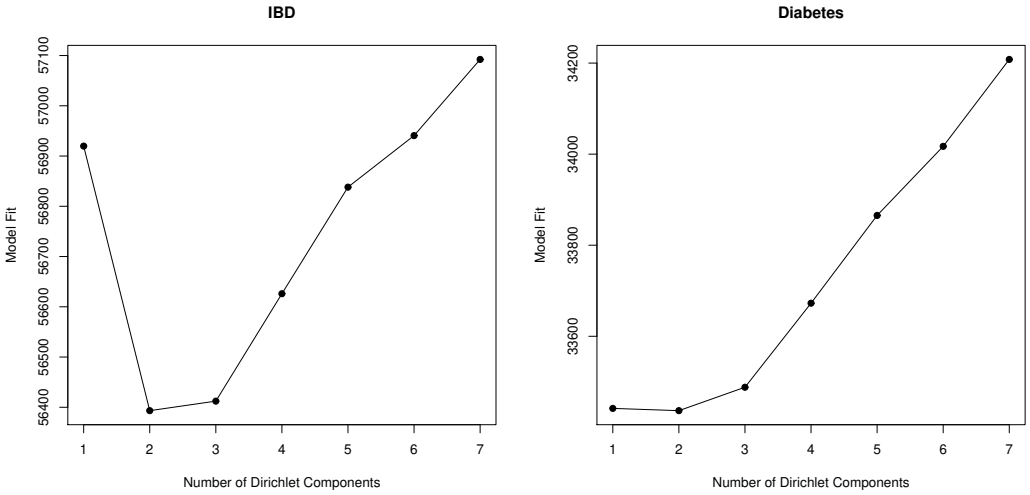

FIGURE B.10: Model fit for DMM with different number of clusters.

Appendix C

Supplementary Material for Chapter 5

C.1 Proof of Lemma 1 and Lemma 2

For Lemma 1, let $\mathcal{T}^{(K)}$ and $\tilde{\mathcal{T}}^{(K)}$ be the underlying partition tree of $\Omega_{\boldsymbol{X}}$ generated when generating $(F_0^{(K)}, F_1^{(K)})$ and $(\tilde{F}_0^{(K)}, \tilde{F}_1^{(K)})$. For any $E \subset \mathscr{T}^{(K)}$, we have $\Pr(\mathcal{T}^{(K)} \in E) = \Pr(\tilde{\mathcal{T}}^{(K)} \in E)$. When $\mathcal{T}^{(K)} = \tilde{\mathcal{T}}^{(K)}$, $(F_0^{(K)}, F_1^{(K)}) \mid A = (\tilde{F}_0^{(K)}, \tilde{F}_1^{(K)}) \mid A$ for $A \in \mathcal{L}(\mathcal{T}^{(K)})$. Note that for $i = 0, 1$, $\tilde{F}_i(\tilde{\mathcal{L}}^{(K)})$ follows a Dirichlet-tree distribution (Dennis, 1991) with parameter $\tilde{\mathcal{T}}^{(K)}$ and $\boldsymbol{\alpha}$ with the following density function

$$p(\boldsymbol{x}) \propto \prod_{A \in \tilde{\mathcal{L}}(\mathcal{T}^{(K)})} x_A^{\alpha(A)-1} \prod_{B \in \tilde{\mathcal{I}}(\mathcal{T}^{(K)})} x_B^{\alpha(B)-\alpha(B_l)-\alpha(B_r)},$$

where $\boldsymbol{x} = \tilde{F}_i(\tilde{\mathcal{L}}^{(K)})$, $x_A = \tilde{F}_i(A)$, $x_B = \sum_{A \in \tilde{\mathcal{T}}_B^{(K)}} x_A$, $\tilde{\mathcal{T}}_B^{(K)}$ is the subtree of $\tilde{\mathcal{T}}^{(K)}$ with B as the root. If $\alpha(A) = \nu G(A)$ for $A \in \mathscr{N}^{(K)}$,

$$\tilde{F}_i(\tilde{\mathcal{L}}^{(K)}) \sim \mathrm{Dir}(b_1, b_2, \ldots, b_{|\tilde{\mathcal{L}}^{(K)}|}) \stackrel{d}{=} F_i(\mathcal{L}^{(K)}),$$

where $b_j = F_i(A_j)$, $A_j \in \mathcal{L}^{(K)}$. □

Lemma 2 is essentially the Theorem 4 in (Ma and Wong, 2011). □

C.2 Some theoretical properties of the nonparametric CDR model

In this section, we provide details on the theoretical properties of the proposed CRD model (without incorporating an outcome model). Specifically, we study the limiting behaviors of the posterior $\pi(\mathcal{P} \mid \mathcal{X}_n, \mathcal{Z}_n)$ as $n \to \infty$. To this end, we introduce some additional notations and list some important notations introduced in Section 5.2. For $\mathcal{T}$, recall that $\mathcal{L}(\mathcal{T})$ and $\mathcal{I}(\mathcal{T})$ denote the set of leaves and internal nodes of $\mathcal{T}$. We write $A \in \mathcal{T}$ if $A \in \mathcal{L}(\mathcal{T}) \cup \mathcal{I}(\mathcal{T})$ Let $n_g = |\boldsymbol{x}_g|$ and $n_g(A) = |\boldsymbol{x}_g(A)|$ for $A \in \mathcal{T}$ and $g = 0, 1$. We denote the covariate of the samples in the two treatment groups as $\mathcal{X}_{0n} = \{\boldsymbol{x}_{01}, \boldsymbol{x}_{02}, \ldots, \boldsymbol{x}_{0n_0}\}$ and $\mathcal{X}_{1n} = \{\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n_1}\}$ where $n = n_0 + n_1$. Let $\mathcal{X}_n = \{\mathcal{X}_{0n}, \mathcal{X}_{1n}\}$ denote the covariate on all the n samples. In this section, we assume that $\mathcal{X}_n$ incorporates the treatment information $\mathcal{Z}_n$. Therefore, we simply write "$\cdot \mid \mathcal{X}_n$" for "$\cdot \mid \mathcal{X}_n, \mathcal{Z}_n$". Given a tree $\mathcal{T}$, we let $\boldsymbol{x}_g(A) = \{\boldsymbol{x}_{gi} : \boldsymbol{x}_{gi} \in A, i = 1, \ldots, n_g\}$ for $g = 0, 1$ be the samples in the two treatment groups that fall in $A \in \mathcal{T}$, $\mathcal{X}_n(A) = \{\boldsymbol{x}_0(A), \boldsymbol{x}_1(A)\}$.

Recall that $\mathscr{T}^{(K)}$ denotes the set of trees with maximum partition level K under partition rule $\mathcal{R}$, $\mathscr{N}^{(K)} = \{A : A \in \bigcup_{\mathcal{T}^{(K)} \in \mathscr{T}^{(K)}} \mathcal{L}(\mathcal{T}^{(K)})\}$ denotes all reachable subsets of $\Omega_{\boldsymbol{X}}$ by trees up to level K. For $\mathcal{T} \in \mathscr{T}^{(K)}$, we let $\iota(A)$ denote the level of A for $A \in \mathcal{T}$. For $\mathcal{T}, \mathcal{T}' \in \mathscr{T}^{(K)}$, we write $\mathcal{T} \preceq \mathcal{T}'$ if $\mathcal{T}$ is a subtree of $\mathcal{T}'$ and $\mathcal{T} \prec \mathcal{T}'$ if in addition $\mathcal{T} \neq \mathcal{T}'$. For $A \in \mathcal{L}(\mathcal{T})$, let $\mathcal{A}_D^{\mathcal{T}'} = \{A_d \in \mathcal{L}(\mathcal{T}') : A_d \subset A\}$ be the set of all descendant leaves of A in $\mathcal{T}'$, and let $\mathcal{A}_D = \bigcup_{\mathcal{T} \prec \mathcal{T}'} \{A_d \in \mathcal{L}(\mathcal{T}') : A_d \subset A\}$ be the set of all potential descendants of A that are reachable by trees up to level K.

Given $\mathcal{T} \in \mathscr{T}^{(K)}$ and a probability measure G on $\Omega_{\boldsymbol{X}}$, let

$$G_{|\mathcal{T}}(B) = \sum_{A \in \mathcal{L}(\mathcal{T})} G(A) \frac{\mu(A \cap B)}{\mu(A)},$$

for any Borel set $B \subset \Omega_{\boldsymbol{X}}$. $G_{|\mathcal{T}}$ can be viewed as an approximation to G with respect to the partition induced by the tree $\mathcal{T}$. If G has a density function g, $G_{|\mathcal{T}}$ also has

a density function:

$$g_{|\mathcal{T}}(\boldsymbol{x}) = \sum_{A \in \mathcal{L}(\mathcal{T})} \mathbb{1}(\boldsymbol{x} \in A) \frac{G(A)}{\mu(A)}.$$

Moreover, for $A \in \mathcal{T}$, let G_A and g_A denote the conditional distribution and the conditional density of G on A, respectively. We use $H(G)$ to denote the entropy of G and $D_{\mathrm{KL}}(G_1 \parallel G_2)$ the Kullback-Leibler divergence from G_2 to G_1 for two probability measures G_1 and G_2. Without loss of generality, let $\mu(\Omega_{\boldsymbol{X}}) = 1$, we have

Lemma 4. *Let G be a probability measure on $\Omega_{\boldsymbol{X}}$ with density g, U the uniform distribution on $\Omega_{\boldsymbol{X}}$. Then for $\mathcal{T} \in \mathscr{T}^{(K)}$,*

1. *(i). $D_{KL}(G_{|\mathcal{T}} \parallel U) = \sum_{A \in \mathcal{L}(\mathcal{T})} G(A) \log \frac{G(A)}{\mu(A)}$;*

2. *(ii). $D_{KL}(G \parallel G_{|\mathcal{T}}) = -H(G) - D_{KL}(G_{|\mathcal{T}} \parallel U)$.*

Proof. (i) is trivial. For (ii),

$$\begin{aligned} D_{\mathrm{KL}}(G \parallel G_{|\mathcal{T}}) &= \sum_{A \in \mathcal{L}(\mathcal{T})} \int_{\boldsymbol{x} \in A} g(\boldsymbol{x}) \log \frac{g(\boldsymbol{x})\mu(A)}{G(A)} d\mu(\boldsymbol{x}) \\ &= \int_{\boldsymbol{x} \in A} g(\boldsymbol{x}) \log g(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \sum_{A \in \mathcal{L}(\mathcal{T})} G(A) \log \frac{G(A)}{\mu(A)} \\ &= -H(G) - D_{\mathrm{KL}}(G_{|\mathcal{T}} \parallel U). \end{aligned}$$

□

The next Lemma studies the limiting behavior of the likelihood given a specific $\mathcal{T}$, which is useful for studying the limiting behavior of the posterior of $\mathcal{P}$. Based on this Lemma, the KL-divergence between the true covariate distributions and the tree-based approximations to these distributions plays an important role.

Lemma 5. *For $\mathcal{T} \in \mathscr{T}^{(K)}$,*

$$\begin{aligned}\frac{l(\mathcal{X}_n \mid \mathcal{T})}{n} &\xrightarrow{p} \zeta_0 D_{KL}(F_{0|\mathcal{T}} \parallel U) + \zeta_1 D_{KL}(F_{1|\mathcal{T}} \parallel U) - D_{KL}(F_{|\mathcal{T}} \parallel U) - H(F) \\ &= D_{KL}(F \parallel F_{|\mathcal{T}}) - \left[\zeta_0 D_{KL}(F_0 \parallel F_{0|\mathcal{T}}) + \zeta_1 D_{KL}(F_1 \parallel F_{1|\mathcal{T}})\right] + C(F_0, F_1),\end{aligned}$$

as $n_0 \to \infty, n_1 \to \infty$ and $n_0/(n_0+n_1) \to \zeta_0 \in (0,1)$, where with respect to $\mathcal{T}$, $C(F_0, F_1) = -\left[\zeta_0 H(F_0) + \zeta_1 H(F_1)\right]$ is a constant.

Proof. Let $\zeta_1 = 1 - \zeta_0$, $F = \zeta_0 F_0 + \zeta_1 F_1$ and $f = \zeta_0 f_0 + \zeta_1 f_1$. Let U be the uniform distribution on $\Omega_{\boldsymbol{X}}$, since $\mu(\Omega_{\boldsymbol{X}}) = 1$, the likelihood of the uniform distribution on $\mathcal{X}_n$ can be decomposed along $\mathcal{T}$ as

$$1 = u(\mathcal{X}_n) = \prod_{A \in \mathcal{N}(\mathcal{T})} V_A(\mathcal{X}_n(A) \mid \mathcal{T}) \times \prod_{A \in \mathcal{L}(\mathcal{T})} W(\mathcal{X}_n(A) \mid A),$$

where

$$\begin{aligned}V_A(\mathcal{X}_n(A) \mid \mathcal{T}) &= \mu(A_l \mid A)^{n(A_l)} \mu(A_r \mid A)^{n(A_r)}, \text{ for } A \in \mathcal{N}(\mathcal{T}) \\ W(\mathcal{X}_n(A) \mid A) &= u(\mathcal{X}_n(A) \mid A) = \mu(A)^{-n(A)}, \text{ for } A \in \mathcal{L}(\mathcal{T}).\end{aligned}$$

Similarly, the marginal likelihood of $\mathcal{T}$ has the decomposition

$$L(\mathcal{X}_n \mid \mathcal{T}) = \prod_{A \in \mathcal{N}(\mathcal{T})} L_A(\mathcal{X}_n(A) \mid \mathcal{T}) \times \prod_{A \in \mathcal{L}(\mathcal{T})} \Phi(\mathcal{X}_n(A) \mid A),$$

where

$$\begin{aligned}L_A(\mathcal{X}_n(A) \mid \mathcal{T}) = &\frac{B(n_0(A_l) + \alpha_0(A_l), n_0(A_r) + \alpha_0(A_r))}{B(\alpha_0(A_l), \alpha_0(A_r))} \\ &\times \frac{B(n_1(A_l) + \alpha_1(A_l), n_1(A_r) + \alpha_1(A_r))}{B(\alpha_1(A_l), \alpha_1(A_r))},\end{aligned}$$

$\Phi(\mathcal{X}_n(A) \mid A)$ is the marginal likelihood of the OPT process on A. Let

$$\begin{aligned}v_A(\mathcal{X}_n(A) \mid \mathcal{T}) &= \log V_A(\mathcal{X}_n(A) \mid \mathcal{T}), \quad w(\mathcal{X}_n(A) \mid A) = \log W(\mathcal{X}_n(A) \mid A) \\ l_A(\mathcal{X}_n(A) \mid \mathcal{T}) &= \log L_A(\mathcal{X}_n(A) \mid \mathcal{T}), \quad \phi(\mathcal{X}_n(A) \mid A) = \log \Phi(\mathcal{X}_n(A) \mid A).\end{aligned}$$

With Stirling's approximation for the Gamma function, we have

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \simeq \sqrt{2\pi}\frac{x^{x-1/2}y^{y-1/2}}{(x+y)^{x+y-1/2}}$$

for large x and y. Following the strategy in (Holmes et al., 2015) and (Soriano and Ma, 2017), for $A \in \mathcal{N}(\mathcal{T})$, we can approximate the ratio

$$\frac{L_A(\mathcal{X}_n(A) \mid \mathcal{T})}{V_A(\mathcal{X}_n(A) \mid \mathcal{T})} \simeq C(\boldsymbol{\alpha}) \prod_{g=0,1} \sqrt{\frac{n_g(A)}{n_g(A_l)n_g(A_r)}} \frac{\hat{F}_g(A_l \mid A)^{n_g(A_l)+\alpha_g(A_l)} \hat{F}_g(A_r \mid A)^{n_g(A_r)+\alpha_g(A_r)}}{\mu(A_l \mid A)^{n_g(A_l)} \mu(A_r \mid A)^{n_g(A_r)}},$$

where $C(\boldsymbol{\alpha})$ is a constant with respect to n, $\hat{F}_g(A_l \mid A) = n_g(A_l)/n_g(A)$ and $\hat{F}_g(A_r \mid A) = n_g(A_r)/n_g(A)$ for $g = 0, 1$. Therefore,

$$\frac{l_A(\mathcal{X}_n(A) \mid A) - v_A(\mathcal{X}_n(A) \mid A)}{n} \xrightarrow{p} \sum_{g=0,1} \zeta_g \left[F_g(A_l) \log \frac{F_g(A_l \mid A)}{\mu(A_l \mid A)} + F_g(A_r) \log \frac{F_g(A_r \mid A)}{\mu(A_r \mid A)} \right].$$

It follows that

$$\begin{aligned}
&\sum_{A \in \mathcal{N}(\mathcal{T})} \frac{l_A(\mathcal{X}_n(A) \mid A) - v_A(\mathcal{X}_n(A) \mid A)}{n} \\
&\xrightarrow{p} \sum_{A \in \mathcal{N}(\mathcal{T})} \sum_{g=0,1} \zeta_g \left[F_g(A_l) \log \frac{F_g(A_l \mid A)}{\mu(A_l \mid A)} + F_g(A_r) \log \frac{F_g(A_r \mid A)}{\mu(A_r \mid A)} \right] \\
&= \sum_{g=0,1} \zeta_g \sum_{A \in \mathcal{N}(\mathcal{T})} \left[\sum_{\{A' \in \mathcal{L}(\mathcal{T}) : A' \subset A_l\}} F_g(A') \log \frac{F_g(A_l \mid A)}{\mu(A_l \mid A)} + \sum_{\{A' \in \mathcal{L}(\mathcal{T}) : A' \subset A_r\}} F_g(A') \log \frac{F_g(A_r \mid A)}{\mu(A_r \mid A)} \right] \\
&= \sum_{g=0,1} \zeta_g \sum_{A' \in \mathcal{L}(\mathcal{T})} \left[F_g(A') \sum_{\{A \in \mathcal{N}(\mathcal{T}) : A \supset A'\}} \left(\mathbb{1}_{(A' \subset A_l)} \log \frac{F_g(A_l \mid A)}{\mu(A_l \mid A)} + \mathbb{1}_{(A' \subset A_r)} \log \frac{F_g(A_r \mid A)}{\mu(A_r \mid A)} \right) \right] \\
&= \sum_{g=0,1} \zeta_g \sum_{A' \in \mathcal{L}(\mathcal{T})} F_g(A') \log \frac{F_g(A')}{\mu(A')} \\
&= \sum_{g=0,1} \zeta_g D_{\mathrm{KL}}(F_{g|\mathcal{T}} \parallel U).
\end{aligned}$$

By Theorem 4 in Wong et al. (2010), the OPT process on each $A \in \mathcal{L}(\mathcal{T})$ is weakly consistent, thus

$$\Phi(\mathcal{X}_n(A) \mid A) \to f(\mathcal{X}_n(A) \mid A), \ a.e.\ F_A^\infty.$$

If follows that

$$\begin{aligned}\frac{\phi(\mathcal{X}_n(A) \mid A) - w(\mathcal{X}_n(A) \mid A)}{n} &\xrightarrow{p} \frac{n_A}{n} \cdot \left[\frac{1}{n_A} \sum_{\boldsymbol{x} \in \mathcal{X}_n(A)} \log \frac{f(\boldsymbol{x} \mid A)}{u(\boldsymbol{x} \mid A)} \right] \\ &\xrightarrow{p} F(A) D_{\mathrm{KL}}(F_A \parallel U_A).\end{aligned}$$

Note that

$$\begin{aligned}D_{\mathrm{KL}}(F_A \parallel U_A) &= \int_{\boldsymbol{x} \in A} f(\boldsymbol{x} \mid A) \log \frac{f(\boldsymbol{x} \mid A)}{u(\boldsymbol{x} \mid A)} d\mu(\boldsymbol{x}) \\ &= \int_{\boldsymbol{x} \in A} \frac{f(\boldsymbol{x})}{F(A)} \log \frac{f(\boldsymbol{x}) \mu(A)}{F(A)} d\mu(\boldsymbol{x}) \\ &= \frac{1}{F(A)} \int_{\boldsymbol{x} \in A} f(\boldsymbol{x}) \log f(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \log \frac{F(A)}{\mu(A)}.\end{aligned}$$

We have

$$\sum_{A \in \mathcal{L}(\mathcal{T})} \frac{\phi(\mathcal{X}_n(A) \mid A) - w(\mathcal{X}_n(A) \mid A)}{n} \xrightarrow{p} -H(F) - D_{\mathrm{KL}}(F_{|\mathcal{T}} \parallel U).$$

Put the two pieces together and apply Lemma 4, we have

$$\begin{aligned}\frac{l(\mathcal{X}_n \mid \mathcal{T})}{n} &= \frac{l(\mathcal{X}_n \mid \mathcal{T}) - \log u(\mathcal{X}_n)}{n} \\ &= \sum_{A \in \mathcal{N}(\mathcal{T})} \frac{l_A(\mathcal{X}_n(A) \mid A) - v_A(\mathcal{X}_n(A) \mid A)}{n} + \sum_{A \in \mathcal{L}(\mathcal{T})} \frac{\phi(\mathcal{X}_n(A) \mid A) - w(\mathcal{X}_n(A) \mid A)}{n} \\ &\xrightarrow{p} \zeta_0 D_{\mathrm{KL}}(F_{0|\mathcal{T}} \parallel U) + \zeta_1 D_{\mathrm{KL}}(F_{1|\mathcal{T}} \parallel U) - D_{\mathrm{KL}}(F_{|\mathcal{T}} \parallel U) - H(F) \\ &= -\zeta_0 H(F_0) - \zeta_1 H(F_1) - \zeta_0 D_{\mathrm{KL}}(F_0 \parallel F_{0|\mathcal{T}}) - \zeta_1 D_{\mathrm{KL}}(F_1 \parallel F_{1|\mathcal{T}}) + D_{\mathrm{KL}}(F \parallel F_{|\mathcal{T}}).\end{aligned}$$

□

The following Lemma characterizes the trees with positive probability mass in the limiting posterior distribution. Specifically, such trees should not contain leaves in which the conditional covariate distributions under the two treatment groups differ (up to a maximum resolution), and should not contain internal nodes in which the conditional covariate distributions under the two treatment groups are the same (up to a maximum resolution). Or put it another way, the procedure for generating the partition stops when it should stop, and does not stop when it should not.

Lemma 6. *Suppose that* $\mathcal{T} \in \mathscr{T}^{(K)}$. *Then* $\Pi(\mathcal{T} \mid \mathcal{X}_n) \to 0$ *as* $n \to \infty$ *if either one of the following conditions holds:*

(i). $\exists A \in \mathcal{L}(\mathcal{T})$ *with* $\iota(A) < K$ *such that* $F_0(A_d \mid A) \neq F_1(A_d \mid A)$ *for some* $A_d \in \mathcal{A}_D$*;*

(ii). $\exists A \in \mathcal{I}(\mathcal{T})$ *such that* $F_0(A_d \mid A) = F_1(A_d \mid A)$ *for all* $A_d \in \mathcal{T}$ *and* $A_d \subset A$.

Proof. For condition (i), by the definition of $\mathcal{A}_D$, $\exists \tilde{\mathcal{T}}$ such that (i) $\mathcal{T} \prec \tilde{\mathcal{T}}$, (ii) $A' \in \mathcal{L}(\tilde{\mathcal{T}})$ if $A' \in \mathcal{L}(\mathcal{T})$ and $A' \neq A$ and (iii) $A_d \in \mathcal{L}(\tilde{\mathcal{T}})$. Note that $\tilde{\mathcal{T}}$ can be obtained by further partition A (and only A) $\in \mathcal{L}(\mathcal{T})$. By Lemma 5,

$$\begin{aligned}
\mathscr{L}_{\mathcal{T}} &= \lim_{n\to\infty} \frac{l(\mathcal{X}_n \mid \mathcal{T})}{n} \\
&= \Delta(\mathcal{X}_n \mid \mathcal{T}\backslash A) + \zeta_0 F_0(A) \log \frac{F_0(A)}{\mu(A)} + \zeta_1 F_1(A) \log \frac{F_1(A)}{\mu(A)} - F(A) \log \frac{F(A)}{\mu(A)} \\
\mathscr{L}_{\tilde{\mathcal{T}}} &= \lim_{n\to\infty} \frac{l(\mathcal{X}_n \mid \tilde{\mathcal{T}})}{n} \\
&= \Delta(\mathcal{X}_n \mid \mathcal{T}\backslash A) + \zeta_0 \cdot \mathcal{G}(F_0, \mathcal{A}_D^{\tilde{\mathcal{T}}}) + \zeta_1 \cdot \mathcal{G}(F_1, \mathcal{A}_D^{\tilde{\mathcal{T}}}) - \mathcal{G}(F, \mathcal{A}_D^{\tilde{\mathcal{T}}}),
\end{aligned}$$

where $\Delta(\mathcal{X}_n \mid \mathcal{T}\backslash A)$ denotes the limit of the log-likelihood at nodes of $\mathcal{T}$ except for

A over n, and for a probability measure G on $\Omega_{\boldsymbol{X}}$,

$$\begin{aligned}
\mathcal{G}(G, \mathcal{A}_D^{\tilde{\mathcal{T}}}) &= \sum_{A_j} G(A_j) \log \frac{G(A_j)}{\mu(A_j)} \\
&= \sum_{A_j} G(A_j) \log \frac{G(A_j \mid A) G(A)}{\mu(A_j \mid A) \mu(A)} \\
&= \sum_{A_j} G(A_j) \log \frac{G(A)}{\mu(A)} + \sum_{A_j} G(A_j) \log \frac{G(A_j \mid A)}{\mu(A_j \mid A)} \\
&= G(A) \log \frac{G(A)}{\mu(A)} + \sum_{A_j} G(A_j) \log \frac{G(A_j)}{\mu(A_j \mid A) G(A)},
\end{aligned}$$

where the summations are taken over $A_j \in \mathcal{A}_D^{\tilde{\mathcal{T}}}$. It follows that

$$\begin{aligned}
\mathscr{L}_{\tilde{\mathcal{T}}} - \mathscr{L}_{\mathcal{T}} &= \sum_{g=0,1} \sum_{A_j} \zeta_g F_g(A_j) \log \frac{F_g(A_j)}{\mu(A_j \mid A) F_g(A)} - \sum_{A_j} F(A_j) \log \frac{F(A_j)}{\mu(A_j \mid A) F(A)} \\
&= \sum_{A_j} \Bigg[\sum_{g=0,1} \zeta_g F_g(A_j) \log \frac{F_g(A_j)}{\mu(A_j \mid A) F_g(A)} \\
&\quad - \left(\sum_{g=0,1} \zeta_g F_g(A_j) \right) \log \frac{\sum_{g=0,1} \zeta_g F_g(A_j)}{\sum_{g=0,1} \zeta_g \mu(A_j \mid A) F_g(A)} \Bigg] \\
&\geqslant 0,
\end{aligned}$$

with equality if and only if for $A_j \in \mathcal{A}_D^{\tilde{\mathcal{T}}}$,

$$\frac{\zeta_0 F_0(A_j)}{\zeta_0 \mu(A_j \mid A) F_0(A)} = \frac{\zeta_1 F_1(A_j)}{\zeta_1 \mu(A_j \mid A) F_1(A)} \quad \Leftrightarrow \quad F_0(A_j \mid A) = F_1(A_j \mid A).$$

Since $F_0(A_d \mid A) \neq F_1(A_d \mid A)$, $\mathscr{L}_{\tilde{\mathcal{T}}} - \mathscr{L}_{\mathcal{T}} > 0$. Thus

$$\frac{\Pi(\mathcal{T} \mid \mathcal{X}_n)}{\Pi(\tilde{\mathcal{T}} \mid \mathcal{X}_n)} = \frac{\Pi(\mathcal{T}) L(\mathcal{X}_n \mid \mathcal{T})}{\Pi(\tilde{\mathcal{T}}) L(\mathcal{X}_n \mid \tilde{\mathcal{T}})} \to 0,$$

which implies that $\Pi(\mathcal{T} \mid \mathcal{X}_n) \to 0$.

For condition (ii), let $\mathcal{T}' \prec \mathcal{T}$ be the tree obtained by pruning the subtree of $\mathcal{T}$ with A_l and A_r as the root in $\mathcal{T}$. We have

$$L(\mathcal{X}_n \mid \mathcal{T}') \geqslant \pi(\mathcal{T} \mid \mathcal{T}') M(\mathcal{X}_n \mid \mathcal{T}),$$

where $\pi(\mathcal{T} \mid \mathcal{T}')$ is the conditional prior probability of $\mathcal{T}$ given $\mathcal{T}'$,

$$M(\mathcal{X}_n \mid \mathcal{T}) = \prod_{B \in \mathcal{N}(\mathcal{T}')} L_B(\mathcal{X}_n(B) \mid \mathcal{T}') \prod_{B \in \mathcal{L}(\mathcal{T})} \Phi(\mathcal{X}_n(B) \mid B)$$

$$\prod_{B \in \mathcal{N}(\mathcal{T}) \backslash \mathcal{N}(\mathcal{T}')} L_B^c(\mathcal{X}_n(B) \mid \mathcal{T}),$$

$L_B(\mathcal{X}_n(B) \mid \mathcal{T})$ and $\Phi(\mathcal{X}_n(B) \mid B)$ are defined in the proof of Lemma 5,

$$L_B^c(\mathcal{X}_n(B) \mid \mathcal{T}) = \frac{B(n_0(B_l) + n_1(B_l) + \omega(B_l), n_0(B_r) + n_1(B_r) + \omega(B_r))}{B(\omega(B_l), \omega(B_r))}.$$

Here $M(\mathcal{X}_n \mid \mathcal{T})$ is the marginal likelihood of $\mathcal{T}$ with the two covariates distributions coupled at all $B \in \mathcal{N}(\mathcal{T}) \backslash \mathcal{N}(\mathcal{T}')$. Note that

$$\frac{M(\mathcal{X}_n \mid \mathcal{T})}{L(\mathcal{X}_n \mid \mathcal{T})} = \prod_{B \in \mathcal{N}(\mathcal{T}) \backslash \mathcal{N}(\mathcal{T}')} \frac{L_B^c(\mathcal{X}_n(B) \mid \mathcal{T})}{L_B(\mathcal{X}_n(B) \mid \mathcal{T})}.$$

Similar to the proof of Lemma 5,

$$\frac{L_B^c(\mathcal{X}_n(B) \mid \mathcal{T})}{L_B(\mathcal{X}_n(B) \mid \mathcal{T})} \simeq C(\boldsymbol{\alpha}, \boldsymbol{\omega}) \sqrt{\frac{n(B)}{n(B_l) n(B_r)}} \prod_{g=0,1} \left[\sqrt{\frac{n_g(B_l) n_g(B_r)}{n_g(B)}} \right.$$

$$\left. \times \frac{\hat{F}(B_l \mid B)^{n(B_l)+\omega(B_l)} \hat{F}(B_r \mid B)^{n(B_r)+\omega(B_r)}}{\hat{F}_g(B_l \mid B)^{n_g(B_l)+\alpha_g(B_l)} \hat{F}_g(B_r \mid B)^{n_g(B_r)+\alpha_g(B_r)}} \right].$$

where $C(\boldsymbol{\alpha}, \boldsymbol{\omega})$ is a constant with respect to n, $\hat{F}_g(B_l \mid B) = n_g(B_l)/n_g(B)$ and $\hat{F}_g(B_r \mid B) = n_g(B_r)/n_g(B)$ for $g = 0, 1$, $\hat{F}(B_l \mid B) = n(B_l)/n(B)$ and $\hat{F}(B_r \mid B) = n(B_r)/n(B)$. Thus

$$l_B^c(\mathcal{X}_n(A) \mid A) - l_B(\mathcal{X}_n(A) \mid A) \simeq \log C(\boldsymbol{\alpha}, \boldsymbol{\omega}) + \frac{1}{2} \log n + \log \Lambda(B),$$

where

$$\log \Lambda(B) := \prod_{g=0,1} \frac{\hat{F}(B_l \mid B)^{n(B_l)} \hat{F}(B_r \mid B)^{n(B_r)}}{\hat{F}_g(B_l \mid B)^{n_g(B_l)} \hat{F}_g(B_r \mid B)^{n_g(B_r)}}$$

is the likelihood ratio for testing composite hypotheses $H_0 : p_0 = p_1$ vs. $H_1 : p_0 \neq p_1 \in (0, 1)$. Under the null, $-2\Lambda(B) \xrightarrow{p} \chi^2_1$ (**?**). Therefore,

$$\frac{l^c_B(\mathcal{X}_n(B) \mid B) - l_B(\mathcal{X}_n(B) \mid B)}{\log n} \xrightarrow{p} \frac{1}{2}.$$

It follows that

$$\begin{aligned}
\frac{L(\mathcal{X}_n \mid \mathcal{T}')}{L(\mathcal{X}_n \mid \mathcal{T})} &\geqslant \pi(\mathcal{T} \mid \mathcal{T}') \frac{M(\mathcal{X}_n \mid \mathcal{T})}{L(\mathcal{X}_n \mid \mathcal{T})} \\
&\xrightarrow{p} \pi(\mathcal{T} \mid \mathcal{T}') \exp\left(\frac{|\mathcal{N}(\mathcal{T})| - |\mathcal{N}(\mathcal{T}')|}{2} \log n \right) \\
&\xrightarrow{p} \infty.
\end{aligned}$$

Thus

$$\frac{\Pi(\mathcal{T} \mid \mathcal{X}_n)}{\Pi(\mathcal{T}' \mid \mathcal{X}_n)} = \frac{\Pi(\mathcal{T}) L(\mathcal{X}_n \mid \mathcal{T})}{\Pi(\mathcal{T}') L(\mathcal{X}_n \mid \mathcal{T}')} \to 0,$$

which implies that $\Pi(\mathcal{T} \mid \mathcal{X}_n) \to 0$. □

Lemma 7, Lemma 8 and Lemma 10 characterize the limit of the posterior on the trees from another perspective. Specifically, these Lemmas profile the limiting behavior of the posterior with respect to the propensity score. We note that Lemma 7 and Lemma 8 are special cases of Lemma 10. We state them separately since they correspond to three common application scenarios.

Let $e(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ be the propensity score function. We first consider the case where the propensity score function induces a sub-stratification structure on $\Omega_{\boldsymbol{X}}$. We call $\mathcal{P}_e$ a propensity score partition of the covariate space if $\mathcal{P}_e = \{E_j \in \mathscr{B}(\mathbb{R}^p) : j \in I_e\}$ such that

1. $\Omega_{\boldsymbol{X}} = \bigcup_{j \in I_e} E_j$;

2. $e(\boldsymbol{x}_1) = e(\boldsymbol{x}_2)$, $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in E_j$, $j \in I_e$;

3. $e(\boldsymbol{x}_1) \neq e(\boldsymbol{x}_2)$, $\forall \boldsymbol{x}_1 \in E_{j_1}$, $\boldsymbol{x}_2 \in E_{j_2}$ and $j_1 \neq j_2$.

where p is the dimension of $\boldsymbol{X}$, $\mathscr{B}(\mathbb{R}^p)$ the Borel algebra on $\mathbb{R}^p$. Let $\mathsf{R}(\mathcal{P}_e)$ be the set of partitions that are refinement of $\mathcal{P}_e$, that is, if $\mathcal{P}_r \in \mathsf{R}(\mathcal{P}_e)$, then $\forall A \in \mathcal{P}_r$, $\exists E_j \in \mathcal{P}_e$ such that $A \subset E_j$. We refer to $\partial\mathcal{P}_e = \bigcup_{j \in I_e} \partial E_j$ as the boundary of $\mathcal{P}_e$.

For $K \geqslant 0$, each $\mathcal{T} \in \mathscr{T}^{(K)}$ induces a partition $\mathcal{L}(\mathcal{T})$ of $\Omega_{\boldsymbol{X}}$. Let $\mathscr{P}^{(K)} = \{\mathcal{L}(\mathcal{T}) : \mathcal{T} \in \mathscr{T}^{(K)}\}$ be the set of all partitions of $\Omega_{\boldsymbol{X}}$ induced by trees with maximum level K under the RDDP prior. Moreover, let $\mathscr{P} = \bigcup_{K \geqslant 0} \mathscr{P}^{(K)}$ denote all partitions that can be generated by the RDDP prior.

Lemma 7. *Suppose that there is a propensity score partition $\mathcal{P}_e$ of $\Omega_{\boldsymbol{X}}$. If $\exists k_0$ such that $\mathcal{P}_e \in \mathscr{P}^{(k_0)}$. Then for $K \geqslant k_0$, $\Pi(\mathcal{L}(\mathcal{T}) \in \mathsf{R}(\mathcal{P}_e) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$.*

Proof. For any $K \geqslant k_0$ and $\mathcal{T} \in \mathscr{T}^{(K)}$, by Lemma 5,

$$\begin{aligned}
\frac{l(\mathcal{X}_n \mid \mathcal{T})}{n} \xrightarrow{p} & \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\int_{\boldsymbol{x} \in A} f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})\mu(A)}{F(A)} d\mu(\boldsymbol{x}) \right. \\
& \left. - \left(\sum_{g=0,1} \zeta_g \int_{\boldsymbol{x} \in A} f_g(\boldsymbol{x}) \log \frac{f_g(\boldsymbol{x})\mu(A)}{F_g(A)} d\mu(\boldsymbol{x}) \right) \right] + C(F_0, F_1) \\
& \leqslant C(F_0, F_1),
\end{aligned}$$

with equality if and only if

$$\begin{aligned}
& \frac{f_0(\boldsymbol{x}) \mathbb{1}_{(\boldsymbol{x} \in A)}}{F_0(A)} = \frac{f_1(\boldsymbol{x}) \mathbb{1}_{(\boldsymbol{x} \in A)}}{F_1(A)} \text{ for } A \in \mathcal{L}(\mathcal{T}), \quad a.e.\ \mu \\
\Leftrightarrow \quad & f_0(\boldsymbol{x} \mid A) = f_1(\boldsymbol{x} \mid A) \text{ for } A \in \mathcal{L}(\mathcal{T}), \quad a.e.\ \mu \\
\Leftrightarrow \quad & e(\boldsymbol{x}) = c_A \text{ for } A \in \mathcal{L}(\mathcal{T}), \quad a.e.\ \mu \\
\Leftrightarrow \quad & \mathcal{L}(\mathcal{T}) \in \mathsf{R}(\mathcal{P}_e).
\end{aligned}$$

□

When the propensity score induces a finite partition that cannot be obtained exactly by the RDDP procedure, or it does not induce a finite partition, we next show that the RDDP procedure will concentrate to partitions that induce "good" approximations to the propensity score function.

Given a partition $\mathcal{P}$, for $\epsilon > 0$, let

$$\mathcal{M}_e(\epsilon; \mathcal{P}) = \left\{ \boldsymbol{x} \in \Omega_{\boldsymbol{X}} : \left| e(\boldsymbol{x}) - \sum_{A \in \mathcal{P}} \frac{\zeta_1 F_1(A)}{\zeta_0 F_0(A) + \zeta_1 F_1(A)} \mathbb{1}_{(\boldsymbol{x} \in A)} \right| < \epsilon \right\}.$$

Let $\mathsf{R}_e(\epsilon)$ be the set of partitions of $\Omega_{\boldsymbol{X}}$ such that if $\mathcal{P} \in \mathsf{R}_e(\epsilon)$, $\mu(\mathcal{M}_e(\epsilon; \mathcal{P})) > 1 - \epsilon$. Therefore, $\mathsf{R}_e(\epsilon)$ defines a set of "good" partitions in terms of the propensity score. If the propensity score induces a finite partition $\mathcal{P}_e$, $\mathsf{R}_e(\epsilon) \neq \varnothing$ since $\mathcal{P}_e \in \mathsf{R}_e(\epsilon)$. Let $r(\boldsymbol{x}) = f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ be the covariate density ratio, we have

$$r(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \frac{\zeta_0 e(\boldsymbol{x})}{\zeta_1[1 - e(\boldsymbol{x})]}, \quad e(\boldsymbol{x}) = \frac{\zeta_1 r(\boldsymbol{x})}{\zeta_0 + \zeta_1 r(\boldsymbol{x})}.$$

Lemma 8. *Suppose that there is a finite propensity score partition $\mathcal{P}_e$ of $\Omega_{\boldsymbol{X}}$ such that $\mu(\partial \mathcal{P}_e) = 0$. If in addition, the covariate density ratio $r(\boldsymbol{x})$ is bounded: $0 < \underline{r} \leqslant r(\boldsymbol{x}) \leqslant \bar{r}$. Then for $\epsilon > 0$, $\exists K_\epsilon$ such that for $K \geqslant K_\epsilon$, $\Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$.*

Proof. For fixed $K \geqslant 0$ and $\mathcal{T} \in \mathscr{T}^{(K)}$, let

$$g(\mathcal{T}) = D_{\mathrm{KL}}(F \parallel F_{|\mathcal{T}}) - \left[\zeta_0 D_{\mathrm{KL}}(F_0 \parallel F_{0|\mathcal{T}}) + \zeta_1 D_{\mathrm{KL}}(F_1 \parallel F_{1|\mathcal{T}})\right] = \sum_{A \in \mathcal{L}(\mathcal{T})} g_A(\mathcal{T})$$

$$g_A(\mathcal{T}) = \int_{\boldsymbol{x} \in A} f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})\mu(A)}{F(A)} d\mu(\boldsymbol{x}) - \left(\sum_{g=0,1} \zeta_g \int_{\boldsymbol{x} \in A} f_g(\boldsymbol{x}) \log \frac{f_g(\boldsymbol{x})\mu(A)}{F_g(A)} d\mu(\boldsymbol{x}) \right)$$

$$\mathsf{T}^{(K)} = \operatorname*{arg\,max}_{\mathcal{T} \in \mathscr{T}^{(K)}} g(\mathcal{T}), \quad \mathsf{M}^{(K)} = \max_{\mathcal{T} \in \mathscr{T}^{(K)}} g(\mathcal{T}).$$

By Lemma 5, as $n \to \infty$,

$$\Pi(\mathcal{T} \in \mathsf{T}^{(K)} \mid \mathcal{X}_n) \to 1.$$

Let $\mathsf{M}^{(K)} = \max_{\mathcal{T}\in\mathscr{T}^{(K)}} g(\mathcal{T})$. Since $\mathscr{T}^{(K)} \subset \mathscr{T}^{(K+1)}$, we have $\mathsf{M}^{(K)} \leqslant \mathsf{M}^{(K+1)} \leqslant 0$.

We first show that as $K \to \infty$, $\mathsf{M}^{(K)} \to 0$. To see this, let $K = p \cdot K'$ for $K' \in \mathbb{N}^+$ and let $\mathcal{T}_K \in \mathscr{T}^{(K)}$ be the tree that cut in the $(l+1)$-th dimension for any window A with $\iota(A) \leqslant K$ and $\iota(A) \equiv l \pmod p$. Note that for $A \in \mathcal{L}(\mathcal{T}_K)$, $\mu(A) = 2^{-K}$ and $\operatorname{diam}(A) = \sqrt{p} \cdot 2^{-K'}$. Let $\mathcal{B}_{\mathcal{T}_K} = \{A \in \mathcal{L}(\mathcal{T}_K) : A \in E_j \text{ for some } j \in I_e\}$, $\mathcal{B}^c_{\mathcal{T}_K} = \mathcal{L}(\mathcal{T}_K)\backslash\mathcal{B}_{\mathcal{T}_K}$. By definition, each set in $\mathcal{B}_{\mathcal{T}}$ is a "balancing block" that allows valid causal inference.

For each $E \in \mathcal{P}_e$, let $\mathcal{B}_{\mathcal{T}_K}(E) = \{A \in \mathcal{L}(\mathcal{T}_K) : A \in E\}$. It can be shown that $\mathcal{O}(E) = \lim_{K\to\infty} \bigcup_{A\in\mathcal{B}_{\mathcal{T}_K}(E)} A$, where $\mathcal{O}(E)$ is the interior of E. To see this, note that $\mathcal{T}_{K+p}$ can be obtained by dividing each $A \in \mathcal{L}(\mathcal{T}_K)$ into 2^p equal-sized cubes with side length $2^{-K'-1}$. Then $\mathcal{B}_{\mathcal{T}_{K+p}}(E)$ can be obtained by taking the union of the smaller cubes if they are completely contained in E. As $K \to \infty$, we keep this process indefinitely. Since $\forall \boldsymbol{x} \in E$, there is a cube $A_{\boldsymbol{x}} \in \mathcal{L}(\mathcal{T}_{p\cdot\tilde{K}})$ with side length $2^{-\tilde{K}}$ such that $\boldsymbol{x} \in A_{\boldsymbol{x}} \subset E$. By the definition of $\mathcal{B}(\mathcal{T}_{p\cdot\tilde{K}})$, $A_{\boldsymbol{x}} \in \mathcal{B}(\mathcal{T}_{p\cdot\tilde{K}})$. Therefore, $\bigcup_{A\in\mathcal{B}_{\mathcal{T}_K}} A \to \bigcup_{j\in I_e} \mathcal{O}(E_j)$. Thus

$$\sum_{A\in\mathcal{B}^c(\mathcal{T}_K)} \mu(A) \to \sum_{j\in I_e} \mu(E_j) - \sum_{j\in I_e} \mu(\mathcal{O}(E_j)) = \mu(\mathcal{P}_e) = 0,$$

as $K \to \infty$. For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in A$, $A \in \mathcal{B}_{\mathcal{T}_K}$, $e(\boldsymbol{x}_1) = e(\boldsymbol{x}_2)$ and $r(\boldsymbol{x}_1) = r(\boldsymbol{x}_2)$. Thus $f_0(\boldsymbol{x})/F_0(A) = f_1(\boldsymbol{x})/F_1(A)$ and $g_A(\mathcal{T}_K) = 0$ for $A \in \mathcal{B}_{\mathcal{T}_K}$. Suppose that $\delta < f_0, f_1 < M$. It follows that,

$$\begin{aligned}\mathsf{M}^{(K)} \geqslant g(\mathcal{T}_K) &= \sum_{A\in\mathcal{B}_{\mathcal{T}_K}} g_A(\mathcal{T}_K) + \sum_{A\in\mathcal{B}^c_{\mathcal{T}_K}} g_A(\mathcal{T}_K) \\ &\geqslant 0 + \left(\delta \log\frac{\delta}{M} - \sum_{g=0,1} \zeta_g M \log\frac{M}{\delta}\right) \sum_{A\in\mathcal{B}^c_{\mathcal{T}_K}} \mu(A) \to 0\end{aligned}$$

as $K \to \infty$.

Given a partition $\mathcal{P}$, similar to the definition of $\mathcal{M}_e(\epsilon;\mathcal{P})$, let

$$\mathcal{M}_r(\epsilon;\mathcal{P}) = \left\{ \boldsymbol{x} \in \Omega_{\boldsymbol{X}} : \left| r(\boldsymbol{x}) - \sum_{A\in\mathcal{P}} \frac{F_1(A)}{F_0(A)} \mathbb{1}_{(\boldsymbol{x}\in A)} \right| < \epsilon \right\}.$$

By the uniform continuity of $g(x) = \zeta_1 x/(\zeta_0 + \zeta_1 x)$ on $[\underline{r}, \bar{r}]$, $\exists \epsilon'$ such that $|g(x_1) - g(x_2)| < \epsilon$ for $|x_1 - x_2| < \epsilon'$. Thus $\mathcal{M}_r(\epsilon';\mathcal{P}) \subset \mathcal{M}_e(\epsilon;\mathcal{P})$. For $\boldsymbol{x} \in A$, let

$$\begin{aligned} h(\boldsymbol{x};A) &= f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})\mu(A)}{F(A)} - \left(\sum_{g=0,1} \zeta_g f_g(\boldsymbol{x}) \log \frac{f_g(\boldsymbol{x})\mu(A)}{F_g(A)} \right) \\ &= (b_0 + b_1) \left(\frac{a_0 + a_1}{b_0 + b_1} \log \frac{a_0 + a_1}{b_0 + b_1} - \sum_{g=0,1} \frac{b_g}{b_0 + b_1} \cdot \frac{a_g}{b_g} \log \frac{a_g}{b_g} \right), \end{aligned}$$

where

$$a_g = \zeta_g f_g(\boldsymbol{x}), \quad b_g = \frac{\zeta_g F_g(A)}{\mu(A)}, \quad \text{for } g = 0, 1.$$

For $\boldsymbol{x} \in A \cap \mathcal{M}_r^c(\epsilon';\mathcal{P})$, by Theorem 1 in (**?**),

$$\begin{aligned} h(\boldsymbol{x};A) &\leqslant -\frac{1}{2}(b_0 + b_1) \times \min\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{(a_0 b_1 - a_1 b_0)^2}{b_0 b_1 (b_0 + b_1)^2} \\ &= -\frac{1}{2} \min\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{a_0^2 b_0}{b_1(b_0 + b_1)} \left(\frac{a_1}{a_0} - \frac{b_1}{b_0} \right)^2 \\ &\leqslant -C\epsilon'^2, \end{aligned}$$

where $C > 0$ is a constant that does not depend on $\boldsymbol{x}$ and A. We note that C exists since $f_0(\boldsymbol{x}), f_1(\boldsymbol{x})$ and $r(\boldsymbol{x})$ are all bounded. Thus for $\mathcal{T} \in \bigcup_{K\geqslant 0} \mathscr{T}^{(K)}$ such that $\mu(\mathcal{M}_r(\epsilon';\mathcal{L}(\mathcal{T}))) \leqslant 1 - \epsilon$, we have

$$g(\mathcal{T}) \leqslant \sum_{A\in\mathcal{L}(\mathcal{T})} \int_{\boldsymbol{x}\in A\cap\mathcal{M}_r^c(\epsilon;\mathcal{L}(\mathcal{T}))} h(\boldsymbol{x};A) d\mu(\boldsymbol{x}) \leqslant -C\epsilon'^2 \mu(\mathcal{M}_r^c(\epsilon;\mathcal{L}(\mathcal{T}))) \leqslant -C\epsilon'^2\epsilon.$$

Since $\mathsf{M}^{(K)} \to 0$ as $K \to \infty$, $\exists K_\epsilon$ such that for $K \geqslant K_\epsilon$, $\mathsf{M}^{(K)} > -C\epsilon'^2\epsilon \geqslant g(\mathcal{T})$. Then for $K \geqslant K_\epsilon$, $\Pi(\mathcal{T} : \mu(\mathcal{M}_r(\epsilon';\mathcal{L}(\mathcal{T}))) \leqslant 1 - \epsilon \mid \mathcal{X}_n) \to 0$ as $n \to \infty$. The result then follows by noticing that $\mathscr{T}^{(K)}$ is finite. □

We next focus on the case where the propensity score function does not necessarily induce a sub-stratification structure on $\Omega_{\boldsymbol{X}}$.

Lemma 9. *Suppose that $r(\boldsymbol{x})$ is uniformly continuous and $0 < \underline{r} \leqslant r(\boldsymbol{x}) \leqslant \bar{r}$. Then $\mathsf{R}_e(\epsilon) \neq \varnothing$ for $\forall \epsilon > 0$.*

Proof. Since $g(x) = \zeta_1 x/(\zeta_0 + \zeta_1 x)$ is uniformly continuous on $[\underline{r}, \bar{r}]$, $\exists \epsilon'$ such that $|g(x_1) - g(x_2)| < \epsilon$ for $|x_1 - x_2| < \epsilon'$.

By the uniform continuity of $r(\boldsymbol{x})$, $\exists \delta_\epsilon$ such that $|r(\boldsymbol{x}_1) - r(\boldsymbol{x}_2)| < \epsilon'$ if $|\boldsymbol{x}_1 - \boldsymbol{x}_2| < \delta_\epsilon$. Let $\mathcal{P}$ be a partition of $\Omega_{\boldsymbol{X}}$ such that $\operatorname{diam}(A) < \delta_\epsilon$. Then

$$\begin{aligned}\left| r(\boldsymbol{x}) - \sum_{A \in \mathcal{P}} \frac{F_1(A)}{F_0(A)} \mathbb{1}_{(\boldsymbol{x} \in A)} \right| &= \left| r(\boldsymbol{x}) - \sum_{A \in \mathcal{P}} \frac{\int_{\boldsymbol{x} \in A} r(\boldsymbol{x}) f_0(\boldsymbol{x}) d\mu(\boldsymbol{x})}{\int_{\boldsymbol{x} \in A} f_0(\boldsymbol{x}) d\mu(\boldsymbol{x})} \mathbb{1}_{(\boldsymbol{x} \in A)} \right| \\ &\leqslant \mathbb{1}_{(\boldsymbol{x} \in A)} \max\left\{ |r(\boldsymbol{x}) - \bar{r}_A|, |r(\boldsymbol{x}) - \underline{r}_A| \right\} \leqslant \epsilon',\end{aligned}$$

where $\bar{r}_A = \max_{\boldsymbol{x} \in A} r(\boldsymbol{x})$, $\underline{r}_A = \min_{\boldsymbol{x} \in A} r(\boldsymbol{x})$. It follows that

$$\left| e(\boldsymbol{x}) - \sum_{A \in \mathcal{P}} \frac{\zeta_1 F_1(A)}{\zeta_0 F_0(A) + \zeta_1 F_1(A)} \mathbb{1}_{(\boldsymbol{x} \in A)} \right| = \left| g(r(\boldsymbol{x})) - g\left(\sum_{A \in \mathcal{P}} \frac{F_1(A)}{F_0(A)} \mathbb{1}_{(\boldsymbol{x} \in A)} \right) \right| < \epsilon.$$

Thus $\mathcal{P} \in \mathsf{R}_e(\epsilon)$. □

Lemma 10. *Suppose that the covariate density ratio $r(\boldsymbol{x})$ is uniformly continuous and bounded such that $0 < \underline{r} \leqslant r(\boldsymbol{x}) \leqslant \bar{r}$. Then for $\epsilon > 0$, $\exists K_\epsilon$ such that for $K \geqslant K_\epsilon$, $\Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$.*

Proof. Define $g(\mathcal{T}), g_A(\mathcal{T}), \mathsf{T}^{(K)}$ and $\mathsf{M}^{(K)}$ as in the proof of Lemma 8. Similar to the proof of Lemma 8, it is sufficient to show that $\mathsf{M}^{(K)} \to 0$ as $K \to \infty$. Let $K = p \cdot K'$ for $K' \in \mathbb{N}^+$ and let $\mathcal{T}_K \in \mathscr{T}^{(K)}$ be the tree defined as in the proof of Lemma 8 such that for $A \in \mathcal{L}(\mathcal{T}_K)$, $\operatorname{diam}(A) = \sqrt{p} \cdot 2^{-K'}$. For $\epsilon > 0$, by the uniform continuity of

$r(\boldsymbol{x})$, there is a δ_ϵ such that for $|r(\boldsymbol{x}_1) - r(\boldsymbol{x}_2)| < \epsilon$ for $|\boldsymbol{x}_1 - \boldsymbol{x}_2| < \delta_\epsilon$. We let K' be chosen such that $\text{diam}(A) < \delta_\epsilon$ for $A \in \mathcal{L}(\mathcal{T}_K)$.

For $A \in \mathcal{L}(\mathcal{T}_K)$, we define $h(\boldsymbol{x}; A)$, a_0, b_0, a_1 and b_1 as in the proof of Lemma 8. By Theorem 1 in (**?**),

$$\begin{aligned} h(\boldsymbol{x}; A) &\geqslant -\frac{1}{2}(b_0 + b_1) \times \max\left\{\frac{b_0}{a_0}, \frac{b_1}{a_1}\right\} \times \frac{(a_0 b_1 - a_1 b_0)^2}{b_0 b_1 (b_0 + b_1)^2} \\ &= -\frac{1}{2}\min\left\{\frac{b_0}{a_0}, \frac{b_1}{a_1}\right\} \times \frac{a_0^2 b_0}{b_1(b_0 + b_1)}\left(\frac{a_1}{a_0} - \frac{b_1}{b_0}\right)^2 \\ &\geqslant -C\epsilon^2, \end{aligned}$$

where C is a global constant that does not depend on $\boldsymbol{x}$ or A. Then

$$g(\mathcal{T}_K) = \sum_{A \in \mathcal{L}(\mathcal{T}_K)} g_A(\mathcal{T}_K) = \sum_{A \in \mathcal{L}(\mathcal{T}_K)} \int_{\boldsymbol{x} \in A} h(\boldsymbol{x}; A) d\mu(\boldsymbol{x}) \geqslant -C\epsilon^2$$

and

$$\mathsf{M}^{(K)} \geqslant g(\mathcal{T}_K) \geqslant -C\epsilon^2.$$

Since ϵ is arbitrary, we have $\mathsf{M}^{(K)} \to 0$ as $K \to \infty$. The rest of the proof follows by the same reasoning as in the proof for Lemma 8. □

C.3 Proof of Theorem 1 and Theorem 2

For Theorem 2, let

$$\hat{\tau}_e = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{Z_i Y_i}{e(\boldsymbol{X}_i)} - \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i}{1 - e(\boldsymbol{X}_i)}\right).$$

We have $\hat{\tau}_e \xrightarrow{p} \tau$. For $\underline{e}/2 > \epsilon > 0$, let

$$\hat{\tau}_{\mathcal{L}(\mathcal{T})} - \hat{\tau}_e = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{n(A)}{n_1(A)} - \frac{1}{e(\boldsymbol{X}_i)} \right) Z_i Y_i \mathbb{1}_{(\boldsymbol{X}_i \in A)} \right.$$

$$\left. - \sum_{i=1}^{n} \left(\frac{n(A)}{n_0(A)} - \frac{1}{1 - e(\boldsymbol{X}_i)} \right) (1 - Z_i) Y_i \mathbb{1}_{(\boldsymbol{X}_i \in A)} \right]$$

$$= (\Delta_{11} + \Delta_{12} + \Delta_{13}) - (\Delta_{01} + \Delta_{02} + \Delta_{13}),$$

where

$$\Delta_{11} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{n(A)}{n_1(A)} - \frac{F(A)}{\zeta_1 F_1(A)} \right) Z_i Y_i \mathbb{1}_{(\boldsymbol{X}_i \in A)} \right]$$

$$\Delta_{12} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{F(A)}{\zeta_1 F_1(A)} - \frac{1}{e(\boldsymbol{X}_i)} \right) Z_i Y_i \mathbb{1}_{[\boldsymbol{X}_i \in A \cap \mathcal{M}_e(\epsilon; \mathcal{L}(\mathcal{T}))]} \right]$$

$$\Delta_{13} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{F(A)}{\zeta_1 F_1(A)} - \frac{1}{e(\boldsymbol{X}_i)} \right) Z_i Y_i \mathbb{1}_{[\boldsymbol{X}_i \in A \cap \mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T}))]} \right],$$

and

$$\Delta_{01} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{n(A)}{n_0(A)} - \frac{F(A)}{\zeta_0 F_0(A)} \right) (1 - Z_i) Y_i \mathbb{1}_{(\boldsymbol{X}_i \in A)} \right]$$

$$\Delta_{02} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{F(A)}{\zeta_0 F_0(A)} - \frac{1}{1 - e(\boldsymbol{X}_i)} \right) (1 - Z_i) Y_i \mathbb{1}_{[\boldsymbol{X}_i \in A \cap \mathcal{M}_e(\epsilon; \mathcal{L}(\mathcal{T}))]} \right]$$

$$\Delta_{03} = \frac{1}{n} \sum_{A \in \mathcal{L}(\mathcal{T})} \left[\sum_{i=1}^{n} \left(\frac{F(A)}{\zeta_0 F_0(A)} - \frac{1}{1 - e(\boldsymbol{X}_i)} \right) (1 - Z_i) Y_i \mathbb{1}_{[\boldsymbol{X}_i \in A \cap \mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T}))]} \right].$$

Note that as $n \to \infty$,

$$|\Delta_{11}| \leqslant \sum_{A \in \mathcal{L}(\mathcal{T})} \left| \frac{n(A)}{n_1(A)} - \frac{F(A)}{\zeta_1 F_1(A)} \right| \frac{B n_1(A)}{n}$$

$$\to \sum_{A \in \mathcal{L}(\mathcal{T})} \left| \frac{F(A)}{\zeta_1 F_1(A)} - \frac{F(A)}{\zeta_1 F_1(A)} \right| B F_1(A) = 0,$$

$$|\Delta_{12}| = \sum_{A \in \mathcal{L}(\mathcal{T})} \frac{B\epsilon}{\underline{e}(\underline{e}-\epsilon)} \cdot \frac{n_1(A \cap \mathcal{M}_e(\epsilon; \mathcal{L}(\mathcal{T})))}{n}$$

$$\to \frac{B\epsilon}{\underline{e}(\underline{e}-\epsilon)} \zeta_1 F_1(\mathcal{M}_e(\epsilon; \mathcal{L}(\mathcal{T}))) \leqslant \frac{B\epsilon}{\underline{e}(\underline{e}-\epsilon)},$$

$$|\Delta_{13}| = BC_1 \sum_{A \in \mathcal{L}(\mathcal{T})} \frac{n_1(A \cap \mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T})))}{n}$$

$$\to BC_1 \zeta_1 F_1(\mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T}))) \leqslant BC_1 M \mu(\mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T}))),$$

where $C_1 < \infty$ is a constant such that

$$C_1 \geqslant \max_{\mathcal{T}} \max_{A \in \mathcal{L}(\mathcal{T})} \max_{\boldsymbol{x} \in A} \left| \frac{1}{e(\boldsymbol{x})} - \frac{F(A)}{\zeta_1 F_1(A)} \right|.$$

Note that C_1 exists since $e(\boldsymbol{x})$ and $r(\boldsymbol{x})$ are bounded. Therefore, for $\mathcal{T}$ such that $\mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon)$, as $n \to \infty$,

$$|\Delta_1| \leqslant |\Delta_{11}| + |\Delta_{12}| + |\Delta_{13}| \leqslant \frac{2B\epsilon}{\underline{e}^2} + BC_1 M\epsilon = C_{11}\epsilon.$$

For $\mathcal{T}$ such that $\mathcal{L}(\mathcal{T}) \notin \mathsf{R}_e(\epsilon)$, as $n \to \infty$,

$$|\Delta_1| \leqslant |\Delta_{11}| + |\Delta_{12}| + |\Delta_{13}| \leqslant \frac{2B\epsilon}{\underline{e}^2} + BC_1 M \mu(\mathcal{M}_e^c(\epsilon; \mathcal{L}(\mathcal{T}))) \leqslant \frac{2B}{\underline{e}^2} + BC_1 M = C_{12}.$$

Similarly, we have as $n \to \infty$,

$$|\Delta_0| \leqslant C_{01}\epsilon, \quad \text{for } \mathcal{T} \text{ such that } \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon)$$

$$|\Delta_0| \leqslant C_{02}, \quad \text{for } \mathcal{T} \text{ such that } \mathcal{L}(\mathcal{T}) \notin \mathsf{R}_e(\epsilon).$$

For $K \geqslant 0$,

$$\mathbb{E}[\hat{\tau}_{\mathcal{L}(\mathcal{T})} \mid \mathcal{X}_n, \mathcal{Y}_n] - \tau = \int \hat{\tau}_{\mathcal{L}(\mathcal{T})} d\Pi(\mathcal{T} \mid \mathcal{X}_n) - \tau^{\text{ATE}}$$

$$= \sum_{\mathcal{T} \in \mathscr{T}^{(K)}} \Pi(\mathcal{T} \mid \mathcal{X}_n)[\hat{\tau}_{\mathcal{L}(\mathcal{T})} - \tau^{\text{ATE}}]$$

$$= \sum_{\mathcal{T} \in \mathscr{T}^{(K)}} \Pi(\mathcal{T} \mid \mathcal{X}_n)[\hat{\tau}_{\mathcal{L}(\mathcal{T})} - \hat{\tau}_e] + (\hat{\tau}_e - \tau^{\text{ATE}}).$$

By Lemma 8 or Lemma 10, for every $\epsilon > 0$, $\exists K_\epsilon \geqslant 0$ such that for $K \geqslant K_\epsilon$, $\Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$. Let $K \geqslant K_\epsilon$, it follows that as $n \to \infty$,

$$\begin{aligned}
|\mathbb{E}[\hat{\tau}_{\mathcal{L}(\mathcal{T})} \mid \mathcal{X}_n, \mathcal{Y}_n] - \tau| &\leqslant \sum_{\mathcal{T} \in \mathscr{T}^{(K)}} \Pi(\mathcal{T} \mid \mathcal{X}_n) |\hat{\tau}_{\mathcal{L}(\mathcal{T})} - \hat{\tau}_e| + |\hat{\tau}_e - \tau^{\mathrm{ATE}}| \\
&\leqslant (C_{01} + C_{11})\epsilon \cdot \Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \\
&\quad + (C_{02} + C_{12}) \cdot \Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \notin \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \\
&\to (C_{01} + C_{11})\epsilon.
\end{aligned}$$

The result then follows by noticing that ϵ is arbitrary. □

For Theorem 2, we first prove the following Lemma:

Lemma 11. *Under the assumption of Theorem 2 for each $K \geqslant 0$, $\Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$ for $\epsilon \geqslant \epsilon_K$, where*

$$\epsilon_K = \left(\frac{M^5 \bar{r}}{\delta^5 \underline{r}} \cdot L_r^2 \cdot \frac{p}{4^{\lfloor K/p \rfloor}} \right)^{\frac{1}{3}}.$$

Proof. Define $g(\mathcal{T}), g_A(\mathcal{T}), \mathsf{T}^{(K)}$ and $\mathsf{M}^{(K)}$ as in the proof of Lemma 8. Similar to the proof of Lemma 8, let $K = p \cdot K'$ for $K' \in \mathbb{N}^+$ and let $\mathcal{T}_K \in \mathscr{T}^{(K)}$ be the tree defined as in the proof of Lemma 8 such that for $A \in \mathcal{L}(\mathcal{T}_K)$, $\mathrm{diam}(A) = \sqrt{p} \cdot 2^{-K'}$. By the Lipschitz continuity of $r(\boldsymbol{x})$, $|r(\boldsymbol{x}_1) - r(\boldsymbol{x}_2)| < \mathrm{diam}(A) L_r$ for $|\boldsymbol{x}_1 - \boldsymbol{x}_2| < \mathrm{diam}(A)$.

For $A \in \mathcal{L}(\mathcal{T}_K)$, we define $h(\boldsymbol{x}; A)$, a_0, b_0, a_1 and b_1 as in the proof of Lemma 8. By Theorem 1 in (**?**),

$$\begin{aligned}
h(\boldsymbol{x}; A) &\geqslant -\frac{1}{2}(b_0 + b_1) \times \max\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{(a_0 b_1 - a_1 b_0)^2}{b_0 b_1 (b_0 + b_1)^2} \\
&= -\frac{1}{2} \max\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{a_0^2 b_0}{b_1 (b_0 + b_1)} \left(\frac{a_1}{a_0} - \frac{b_1}{b_0} \right)^2 \\
&\geqslant -C L_r^2 p \cdot 4^{-K'},
\end{aligned}$$

where

$$C = \frac{1}{2}\frac{M^3}{\delta^2 \underline{r}}.$$

Then

$$g(\mathcal{T}_K) = \sum_{A\in\mathcal{L}(\mathcal{T}_K)} g_A(\mathcal{T}_K) = \sum_{A\in\mathcal{L}(\mathcal{T}_K)} \int_{\boldsymbol{x}\in A} h(\boldsymbol{x}; A) d\mu(\boldsymbol{x}) \geqslant -CL_r^2 p \cdot 4^{-K'},$$

and

$$\mathsf{M}^{(K)} \geqslant g(\mathcal{T}_K) \geqslant -CL_r^2 p \cdot 4^{-K'}.$$

Given a partition $\mathcal{P}$, similar to the definition of $\mathcal{M}_e(\epsilon; \mathcal{P})$, let

$$\mathcal{M}_r(\epsilon; \mathcal{P}) = \left\{ \boldsymbol{x} \in \Omega_{\boldsymbol{X}} : \left| r(\boldsymbol{x}) - \sum_{A\in\mathcal{P}} \frac{F_1(A)}{F_0(A)} \mathbb{1}_{(\boldsymbol{x}\in A)} \right| < \epsilon \right\}.$$

For $\epsilon > 0$, by the Lipschitz continuity of $g(x) = \zeta_1 x/(\zeta_0 + \zeta_1 x)$ on $[\underline{r}, \bar{r}]$, $|g(x_1) - g(x_2)| < \epsilon$ for $|x_1 - x_2| < \epsilon' = \epsilon(\zeta_0 + \zeta_1 \underline{r})^2/\zeta_0\zeta_1$. Thus $\mathcal{M}_r(\epsilon'; \mathcal{P}) \subset \mathcal{M}_e(\epsilon; \mathcal{P})$. For $\boldsymbol{x} \in A$, let

$$\begin{aligned} h(\boldsymbol{x}; A) &= f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})\mu(A)}{F(A)} - \left(\sum_{g=0,1} \zeta_g f_g(\boldsymbol{x}) \log \frac{f_g(\boldsymbol{x})\mu(A)}{F_g(A)} \right) \\ &= (b_0 + b_1) \left(\frac{a_0 + a_1}{b_0 + b_1} \log \frac{a_0 + a_1}{b_0 + b_1} - \sum_{g=0,1} \frac{b_g}{b_0 + b_1} \cdot \frac{a_g}{b_g} \log \frac{a_g}{b_g} \right), \end{aligned}$$

where

$$a_g = \zeta_g f_g(\boldsymbol{x}), \quad b_g = \frac{\zeta_g F_g(A)}{\mu(A)}, \quad \text{for } g = 0, 1.$$

For $\boldsymbol{x} \in A \cap \mathcal{M}_r^c(\epsilon'; \mathcal{P})$, by Theorem 1 in (**?**),

$$\begin{aligned} h(\boldsymbol{x}; A) &\leqslant -\frac{1}{2}(b_0 + b_1) \times \min\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{(a_0 b_1 - a_1 b_0)^2}{b_0 b_1 (b_0 + b_1)^2} \\ &= -\frac{1}{2} \min\left\{ \frac{b_0}{a_0}, \frac{b_1}{a_1} \right\} \times \frac{a_0^2 b_0}{b_1(b_0 + b_1)} \left(\frac{a_1}{a_0} - \frac{b_1}{b_0} \right)^2 \\ &< -C'\epsilon'^2, \end{aligned}$$

where

$$C' = \frac{1}{2}\frac{\delta^3}{M^2\bar{r}}.$$

Thus for $\mathcal{T} \in \mathscr{T}^{(K)}$ such that $\mu(\mathcal{M}_r(\epsilon'; \mathcal{L}(\mathcal{T}))) \leqslant 1 - \epsilon$, we have

$$g(\mathcal{T}) \leqslant \sum_{A \in \mathcal{L}(\mathcal{T})} \int_{\boldsymbol{x} \in A \cap \mathcal{M}_r^c(\epsilon; \mathcal{L}(\mathcal{T}))} h(\boldsymbol{x}; A) d\mu(\boldsymbol{x}) < -C'\epsilon'^2 \mu(\mathcal{M}_r^c(\epsilon; \mathcal{L}(\mathcal{T}))) \leqslant -C'\epsilon'^2\epsilon$$

$$= -\frac{1}{2}\frac{\delta^3}{M^2\bar{r}}\frac{(\zeta_0 + \zeta_1\underline{r})^2}{\zeta_0\zeta_1}\epsilon^3.$$

For $\epsilon \geqslant \epsilon_K$, we have

$$g(\mathcal{T}) < -\frac{1}{2}\frac{\delta^3}{M^2\bar{r}}\frac{(\zeta_0 + \zeta_1\underline{r})^2}{\zeta_0\zeta_1}\epsilon^3 \leqslant -CL_r^2\frac{p}{4^{K'}} < \mathsf{M}^{(K)}.$$

By Lemma 5, $\Pi(\mathcal{T} : \mathcal{L}(\mathcal{T}) \in \mathsf{R}_e(\epsilon) \mid \mathcal{X}_n) \to 1$ as $n \to \infty$. □

Following Lemma 11, the proof is essentially the same as the proof for Theorem 1.

Bibliography

Anderson, M. J. (2001), "A new method for non-parametric multivariate analysis of variance," *Austral ecology*, 26, 32–46.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011), "Enterotypes of the human gut microbiome," *nature*, 473, 174.

Atchison, J. and Shen, S. M. (1980), "Logistic-normal distributions: Some properties and uses," *Biometrika*, 67, 261–272.

Athey, S. and Imbens, G. (2016), "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

Barbieri, M. M., Berger, J. O., et al. (2004), "Optimal predictive model selection," *The annals of statistics*, 32, 870–897.

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J., and Pericchi, L. (2001), "Objective Bayesian methods for model selection: introduction and comparison," *Lecture Notes-Monograph Series*, pp. 135–207.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), "Classification and regression trees," .

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016), "DADA2: high-resolution sample inference from Illumina amplicon data," *Nature methods*, 13, 581.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017), "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The ISME journal*, 11, 2639.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010), "QIIME allows analysis of high-throughput community sequencing data," *Nature methods*, 7, 335–336.

Chen, H. and Friedman, J. H. (2017), "A new graph-based two-sample test for multivariate and object data," *Journal of the American statistical association*, 112, 397–409.

Chen, J. and Li, H. (2013), "Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis," *The annals of applied statistics*, 7.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART model search," *Journal of the American Statistical Association*, 93, 935–948.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010), "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, 4, 266–298.

Christensen, J. and Ma, L. (2020), "A Bayesian hierarchical model for related densities by using Pólya trees," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Costea, P. I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M. J., Bushman, F. D., De Vos, W. M., Ehrlich, S. D., Fraser, C. M., Hattori, M., et al. (2018), "Enterotypes in the landscape of gut microbial community composition," *Nature microbiology*, 3, 8–16.

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998), "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Transactions on signal processing*, 46, 886–902.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Dahl, D. B. (2006), "Model-based clustering for expression data via a Dirichlet process mixture model," *Bayesian inference for gene expression and proteomics*, 4, 201–218.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., et al. (2014), "Diet rapidly and reproducibly alters the human gut microbiome," *Nature*, 505, 559–563.

Denison, D. G., Mallick, B. K., and Smith, A. F. (1998), "A bayesian cart algorithm," *Biometrika*, 85, 363–377.

Dennis, S. Y. (1991), "On the hyper-Dirichlet type 1 and hyper-Liouville distributions," *Communications in Statistics-Theory and Methods*, 20, 4069–4081.

Dennis, S. Y. (1996), "A Bayesian analysis of tree-structured statistical decision problems," *Journal of statistical planning and inference*, 53, 323–344.

Dethlefsen, L. and Relman, D. A. (2011), "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation," *Proceedings of the National Academy of Sciences*, 108, 4554–4561.

Ding, T. and Schloss, P. D. (2014), "Dynamics and associations of microbial community types across the human body," *Nature*, 509, 357.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, 90, 577–588.

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian variable selection," *Statistica sinica*, pp. 339–373.

Goedert, J. J., Hua, X., Yu, G., and Shi, J. (2014), "Diversity and composition of the adult fecal microbiome associated with history of cesarean birth or appendectomy: analysis of the American Gut Project," *EBioMedicine*, 1, 167–172.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016), "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, 17, 333.

Graf, D., Di Cagno, R., Fåk, F., Flint, H. J., Nyman, M., Saarela, M., and Watzl, B. (2015), "Contribution of diet to the composition of the human gut microbiota," *Microbial ecology in health and disease*, 26, 26164.

Grantham, N. S., Reich, B. J., Borer, E. T., and Gross, K. (2017), "MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments," *arXiv preprint arXiv:1703.07747*.

Grice, E. A. and Segre, J. A. (2012), "The human microbiome: our second genome," *Annual review of genomics and human genetics*, 13, 151–170.

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. (2020), "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects," *Bayesian Analysis*.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Hanson, T. E. (2006), "Inference for mixtures of finite Polya tree models," *Journal of the American Statistical Association*, 101, 1548–1565.

Held, L., Bové, D. S., Gravestock, I., et al. (2015), "Approximate Bayesian model selection with the deviance statistic," *Statistical Science*, 30, 242–257.

Hildebrandt, M. A., Hoffmann, C., Sherrill-Mix, S. A., Keilbaugh, S. A., Hamady, M., Chen, Y.-Y., Knight, R., Ahima, R. S., Bushman, F., and Wu, G. D. (2009), "High-fat diet determines the composition of the murine gut microbiome independently of obesity," *Gastroenterology*, 137, 1716–1724.

Hill, J. L. (2011), "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, 20, 217–240.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hoff, P. D. (2003), "Nonparametric modeling of hierarchically exchangeable data," *University of Washington Statistics Department, Tech. Rep*, 421.

Hoff, P. D. et al. (2006), "Model-based subspace clustering," *Bayesian Analysis*, 1, 321–344.

Hoff, P. D. et al. (2007), "Extending the rank likelihood for semiparametric copula estimation," *The Annals of Applied Statistics*, 1, 265–283.

Holmes, C. C., Caron, F., Griffin, J. E., Stephens, D. A., et al. (2015), "Two-sample Bayesian nonparametric hypothesis testing," *Bayesian Analysis*, 10, 297–320.

Holmes, I., Harris, K., and Quince, C. (2012), "Dirichlet multinomial mixtures: generative models for microbial metagenomics," *PloS one*, 7, e30126.

Imai, K. and Ratkovic, M. (2014), "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 243–263.

Imai, K., King, G., and Stuart, E. A. (2008), "Misunderstandings between experimentalists and observationalists about causal inference," *Journal of the royal statistical society: series A (statistics in society)*, 171, 481–502.

Imai, K., Ratkovic, M., et al. (2013), "Estimating treatment effect heterogeneity in randomized program evaluation," *The Annals of Applied Statistics*, 7, 443–470.

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Jaccard, P. (1912), "The distribution of the flora in the alpine zone. 1," *New phytologist*, 11, 37–50.

Kang, J. D., Schafer, J. L., et al. (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, 22, 523–539.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013), "Gut metagenome in European women with normal, impaired and diabetic glucose control," *Nature*, 498, 99.

Kaufman, L. and Rousseeuw, P. J. (2009), *Finding groups in data: an introduction to cluster analysis*, vol. 344, John Wiley & Sons.

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011), "Bayesian community-wide culture-independent microbial source tracking," *Nature methods*, 8, 761.

Koller, D. and Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., and Ley, R. E. (2013), "A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets," *PLoS computational biology*, 9, e1002863.

Kuntz, T. M. and Gilbert, J. A. (2017), "Introducing the microbiome into precision medicine," *Trends in pharmacological sciences*, 38, 81–91.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012a), "Hypothesis testing and power calculations for taxonomic-based human microbiome data," *PloS one*, 7, e52078.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012b), "Hypothesis testing and power calculations for taxonomic-based human microbiome data," *PloS one*, 7, e52078.

Lakshminarayanan, B., Roy, D., and Teh, Y. W. (2013), "Top-down particle filtering for Bayesian decision trees," in *International Conference on Machine Learning*, pp. 280–288.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 157–224.

Lavine, M. et al. (1992), "Some aspects of Polya tree distributions for statistical modelling," *The annals of statistics*, 20, 1222–1235.

Lavine, M. et al. (1994), "More aspects of Polya tree distributions for statistical modelling," *The Annals of Statistics*, 22, 1161–1176.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2017), "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, pp. 1–11.

Li, Y. and Clyde, M. A. (2015), "Mixtures of g-priors in Generalized Linear Models," *arXiv preprint arXiv:1503.06913*.

Li, Z., Lee, K., Karagas, M. R., Madan, J. C., Hoen, A. G., O?malley, A. J., and Li, H. (2018), "Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data," *Statistics in biosciences*, 10, 587–608.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Liao, S. X. and Zigler, C. M. (2020), "Uncertainty in the design stage of two-stage Bayesian propensity score analysis," *Statistics in Medicine*.

Liu, F., Bayarri, M., Berger, J., et al. (2009), "Modularization in Bayesian analysis, with emphasis on analysis of computer models," *Bayesian Analysis*, 4, 119–150.

Lloyd, S. (1982), "Least squares quantization in PCM," *IEEE transactions on information theory*, 28, 129–137.

Lozupone, C. and Knight, R. (2005), "UniFrac: a new phylogenetic method for comparing microbial communities," *Appl. Environ. Microbiol.*, 71, 8228–8235.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012), "Diversity, stability and resilience of the human gut microbiota," *Nature*, 489, 220–230.

Lunceford, J. K. and Davidian, M. (2004), "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in medicine*, 23, 2937–2960.

Ma, L. and Soriano, J. (2018), "Analysis of distributional variation through graphical multi-scale Beta-Binomial models," *Journal of Computational and Graphical Statistics*, pp. 1–13.

Ma, L. and Wong, W. H. (2011), "Coupling optional Pólya trees and the two sample problem," *Journal of the American Statistical Association*, 106, 1553–1565.

Mao, J. and Ma, L. (2020), "Dirichlet-tree multinomial mixtures for clustering microbiome compositions," *arXiv preprint arXiv:2008.00400*.

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

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." *Psychological methods*, 9, 403.

McDonald, D., Birmingham, A., and Knight, R. (2015a), "Context and the human microbiome," *Microbiome*, 3, 52.

McDonald, D., Hornig, M., Lozupone, C., Debelius, J., Gilbert, J. A., and Knight, R. (2015b), "Towards large-cohort comparative studies to define the factors influencing the gut microbial community structure of ASD patients," *Microbial ecology in health and disease*, 26.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018), "American gut: an open platform for citizen science microbiome research," *mSystems*, 3, e00031–18.

Miller, J. W. and Harrison, M. T. (2013), "A simple example of Dirichlet process mixture inconsistency for the number of components," in *Advances in neural information processing systems*, pp. 199–206.

Mosimann, J. E. (1962), "On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions," *Biometrika*, 49, 65–82.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal sample size for multiple testing: the case of gene expression microarrays," *Journal of the American Statistical Association*, 99, 990–1001.

Müller, P., Parmigiani, G., and Rice, K. (2006), "FDR and Bayesian multiple comparisons rules," .

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9, 249–265.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002), "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856.

Paddock, S. M., Ruggeri, F., Lavine, M., and West, M. (2003), "Randomized Polya tree models for nonparametric Bayesian inference," *Statistica Sinica*, pp. 443–460.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010), "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, 464, 59.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012), "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, 490, 55–60.

Quince, C., Lundin, E. E., Andreasson, A. N., Greco, D., Rafter, J., Talley, N. J., Agreus, L., Andersson, A. F., Engstrand, L., and D'Amato, M. (2013), "The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study," *Gut*, 62, 952–954.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2017), "Bayesian Nonparametric Mixed Effects Models in Microbiome Data Analysis," *arXiv preprint arXiv:1711.01241*.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

Rosenbaum, P. R. (1989), "Optimal matching for observational studies," *Journal of the American Statistical Association*, 84, 1024–1032.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American statistical Association*, 79, 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., et al. (2018), "Environment dominates over host genetics in shaping human gut microbiota," *Nature*, 555, 210.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66, 688.

Rubin, D. B. (1978), "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, pp. 34–58.

Rubin, D. B. (2008a), "Comment: The design and analysis of gold standard randomized experiments," *Journal of the American Statistical Association*, 103, 1350–1353.

Rubin, D. B. (2008b), "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, pp. 808–840.

Scott, J. G. and Berger, J. O. (2010), "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," *The Annals of Statistics*, pp. 2587–2619.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica sinica*, pp. 639–650.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017), "A phylogenetic transform enhances analysis of compositional microbiota data," *Elife*, 6, e21887.

Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A., and Mukherjee, S. (2019), "Bayesian multinomial logistic normal models through marginally latent matrix-t processes," *arXiv preprint arXiv:1903.11695*.

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H., et al. (2017), "Influence of diet on the gut microbiome and implications for human health," *Journal of translational medicine*, 15, 73.

Smith, H. L. (1997), "Matching with multiple controls to estimate treatment effects in observational studies," *Sociological methodology*, 27, 325–353.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006), "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proceedings of the National Academy of Sciences*, 103, 12115–12120.

Soriano, J. and Ma, L. (2017), "Probabilistic multi-resolution scanning for two-sample differences," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 547–572.

Stuart, E. A. (2010), "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25, 1.

Tang, Y. and Nicolae, D. L. (2017), "Mixed Effect Dirichlet-Tree Multinomial for Longitudinal Microbiome Data and Weight Prediction," *arXiv preprint arXiv:1706.06380*.

Tang, Y., Ma, L., Nicolae, D. L., et al. (2018), "A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data," *The Annals of Applied Statistics*, 12, 1–26.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014), "A simple method for estimating interactions between a treatment and a large number of covariates," *Journal of the American Statistical Association*, 109, 1517–1532.

Tsilimigras, M. C. and Fodor, A. A. (2016), "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," *Annals of epidemiology*, 26, 330–335.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006), "An obesity-associated gut microbiome with increased capacity for energy harvest," *nature*, 444, 1027–131.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009), "A core gut microbiome in obese and lean twins," *nature*, 457, 480.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017), "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data," *BMC bioinformatics*, 18, 94.

Wager, S. and Athey, S. (2018), "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wang, T. and Zhao, H. (2017), "A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms," *Biometrics*.

Wellman, M. P. and Henrion, M. (1993), "Explaining 'explaining away'," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 287–292.

Wong, W. H., Ma, L., et al. (2010), "Optional Pólya tree and Bayesian inference," *The Annals of Statistics*, 38, 1433–1459.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011), "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, 334, 105–108.

Wu, Y., Tjelmeland, H., and West, M. (2007), "Bayesian CART: Prior specification and posterior simulation," *Journal of Computational and Graphical Statistics*, 16, 44–66.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013), "A logistic normal multinomial regression model for microbiome compositional data analysis," *Biometrics*, 69, 1053–1063.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques*.

Biography

Jialiang Mao received the B.S. in Mathematics & Applied Mathematics from Chu Kochen Honors College at Zhejiang University. He then moved to Durham, NC to pursue master's and doctoral studies in the Department of Statistical Science at Duke University under the supervision of Professor Li Ma. In May 2016, he received the M.S. in Statistical Statistics from the Department of Statistical Science at Duke University.