BAYESIAN BALANCE REGRESSION AND MEDIATION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

Lu Huang

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____

Hongzhe Li, Ph.D., Perelman Professor in Biostatistics, Epidemiology, and Informatics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Qi Long, Professor of Biostatistics

Pamela Shaw, Associate Professor of Biostatistics

James Lews, Professor of Medicine and Epidemiology

Kyle Bittinger, Assistant Professor of Pediatrics at the Children's Hospital of Philadelphia

BAYESIAN BALANCE REGRESSION AND MEDIATION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

ACKNOWLEDGMENT

There are so many things that I am grateful for and so many people that I'm lucky to have known throughout my training at Penn. Returning to school after almost 4 years of full-time job was not an easy decision nor an easy task. I've experienced so many bumpy rides during class, with collaborations, and in my research. Magically, each time, though sometimes it took longer than expected, I managed to reach the finish line s afely. When looking back, generous help and support from my advisor, my committee, all the faculty in DBEI, my fellow students and administrative personnel are the indispensable factors in my successful journey at Penn.

Without my advisor, Professor Hongzhe Li, who challenged me to think deeper, I would not have a thorough understanding of my research work nor would I have developed the ability to conduct independent research. Without my committee members, I would not have the precious and continuous suggestions on my projects and various aspects of my research across medical and statistical fields. Without the coursework, homework and exams lectured by all the faculty in the department, I would not have grasped a solid foundation of statistical theory, methodology and applications. Without my fellow students and all the social activities organized by them, I would not have recovered quickly from setbacks either in my personal life or my ongoing research. Lastly, without the administrative personnel in GGEB program or in the DBEI department, I would not have transitioned smoothly and kept track of my progress each year.

Lastly, I owe my deepest appreciation to my family. There is so much to say that I don't know where to start. If I could perform some dimension reduction to my words, the resulting sentence would just be "Thank you!"

ABSTRACT

BAYESIAN BALANCE REGRESSION AND MEDIATION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

Lu Huang

Hongzhe Li

In microbiome studies, 16S rRNA sequencing is commonly used to quantify the taxonomic abundance of a microbial community. The resulting data are counts of amplicons. However, the total count is not informative because of the sampling, sample preparation and sequencing processes. These counts are used to obtain estimates of the relative abundance of the taxa, which is compositional with a unit sum constraint. Analysis of compositional data requires special statistical treatment to account for the intrinsic dependence of the components due to this constraint. Balance, defined as the normalized log ratio of the geometric mean of the values for the two groups of components, provides an interesting way of studying microbial community structure, where the two groups represent the beneficial and detrimental taxa, respectively. Such a balance can be used to quantify dysbiosis of the microbial community that is associated with a clinical outcome. However, identification of the outcome-associated balance is challenging. We introduce a Bayesian balance-regression and a Markov Chain Monte Carlo (MCMC) stochastic search algorithm to identify the compositional balance that is associated with the outcome. Specifically, we propose a random walk strategy in MCMC that explores the very large space of all possible balance defined from high dimensional compositional vector. Simulation studies suggest that the algorithm can identify the bacterial taxa that define the outcome-associated balance with a high probability. The effect of the balance on the outcome can be easily inferred from their predictive posterior distribution. We apply the proposed methods to two human microbiome studies and identify the balance of gut microbiome composition that are associated with body mass index and risk of inflammatory bowel disease, respectively.

Microbial compositional balance can also be used to define a mediator to link treatment or environment factor to an outcome. However, for a given study, the balance that mediates the treatment effect on outcome is unknown. We propose a Bayesian balance mediation model and a Markov

chain Monte Carlo (MCMC) method to simultaneously search for such a balance and to make inference on the mediation effects based on the predictive posterior distributions. Based on the proposed model, we show that the mediation effect can be defined in terms of balance effect on the outcome, balance indicator and the effect of treatment on compositional shift. Our simulation results show that the MCMC sampling can effectively identify the balance and provide correct estimate of the direct and mediation effects. We apply the method to a microbiome study aiming to understand the role of gut microbiome in linking vegan diet to several plasma metabolites. Our analysis shows that vegan diet has strong direct effects and the compositional balance identified has a weak to moderate effect on these plasma metabolites, however, the mediation effects of gut microbiome on these metabolites are very small.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Microbiota and human health

Microbiota refers to a collection of small organisms that live as a community. In particular, human microbiota contains all the bacteria, viruses, fungi, archaea in a particular niche on human body, like skin, oral and respiratory tracks, intestines, to name a few. Those small organisms co-live with us and are an integral part of our body. The total number of microbes in human body is estimated to be in the order of 10^{14}, out-numbering the total of human body cells by 10 to 1 (Ley, Peterson, and Gordon, 2006; Savage, 1977; Whitman, Coleman, and Wiebe, 1998). Such a large number of microbes contributes to maintaining human body's normal functions. For example, microbes participate in nutrient as well as drug metabolism and immunomodulation. The immune system of a newborn is first trained by the environmental microbiota encountered after birth. On the other hand, disruption of microbial community can lead to various diseases. For example, *Clostridium difficile* infection (CDI) is due to an increase of *C.diff*, a bacterium that also resides in healthy person. Therapeutic strategies that specifically target human microbes are under fast development and have shown great promise.

Gut microbiota, often referred as gut flora, takes up about 70% of all microbes in human (Ley, Peterson, and Gordon, 2006; Whitman, Coleman, and Wiebe, 1998). The gut microbiota acts as an extra layer against outside pathogens (Hooper, 2009)and has an indispensable role in processing/generating special metabolites, fermantation of fibers and mucosal immunity (Herbrand et al., 2008). Diseases that are reported to be associated with gut microbiota include obesity, Crohn's disease, Irritable Bowel Syndrome (IBS) (Ley et al., 2005a, 2006), etc. In addition to digestive system abnormalities, gut mcirobiota has also been reported to be linked to functions of various organs including brain, known as gut-brain axis (Cryan et al., 2019), liver and pancreas (Sekirov et al., 2010). Several approaches have been successfully developed to treat gut microbiota-related diseases. Transplatation of fecal samples from healthy donors is an effective procedure for treating *Clostridium difficile* infection (Kassam et al., 2013). It has also been reported that gut microbiota is essential in immunotherapy of certain cancer types (Vétizou et al., 2015). One interesting aspect

of gut microbiota is that our food intake can have impact on this community. As such, probiotics or prebiotics are usually recommended when a patient is taking antibiotics in order to re-establish gut microbial community and to reduce antibiotic related diarrhea.

1.2. Current sequencing methods in microbiota studies

Traditionally microbiologists culture a particular microbe and then study its properties in a wet lab. However, the environmental niche where a microbiota community lives might not be culturable in reality. As a result, microbiologists do not have the capability to culture all the microbes in that environment. Microbiome, the total genome from a microbiota sample, is an alternative approach to studying the composition and functions of all microbes from an environmental sample. Microbiome are usually obtained by sequencing and the resulting sequences are then processed by various bioinformatic tools. Compared to the traditional wet lab approach, culture-independent sequencing of microbiome requires less time as it eliminates the time in culturing each microbe. However, the functional analysis of a microbiota sample can only be inferred from the sequencing results, whereas the wet-lab approach is better suited to study the functions of different microbes or a collection of them. Despite the indirect information in functional analyses, most microbiota/microbiome studies are performed using sequencing methods. The resulting data are often call metagenomic data, reflecting the fact that sequencing results contain genomes from multiple microbes in a given community.

Two sequencing strategies are often used in microbiome studies, the target sequencing and shotgun metagenomic sequencing. The target sequencing approach depends on certain genes or genomic regions that are conserved among a large number of microbes and yet have small variations that can differentiate those microbes at different taxonomic levels. For example, 16S ribosomal RNA (rRNA) gene is a highly conserved gene in bacteria kingdom (Caporaso et al., 2010). It contains several variable regions that can map each bacteria species along the phylogenetic tree. This is the most frequently used target gene approach in microbiome studies. In order to profile bacteria in a sample, researchers first amplify this gene using the conserved sequence and then perform high-throughput sequencing to get the nucleotide information in the variable regions. The sequencing reads are first clustered into Operational Taxonomic Unit (OTU), which represents species with 97% as the similarity threshold in clustering analysis. The corresponding bacteria for each OTU

can be compared with reference microbial genome databases (DeSantis et al., 2006; Kim et al., 2012).

In contrast, the shotgun metagenomic sequencing strategy does not rely on sequencing particular genes. It collects all the genetic materials in a sample, randomly shred the genomes and send the entire genome fragments to sequencers. Assigning each sequencing read to a particular bacterial species and calculating its abundance is a great challenge in analyzing shotgun sequencing data. Because a lot of regions are shared among the species that are close to each other on the phylogenetic tree, it is not possible to assign all the sequencing reads from these regions to a particular species. In the last 20 years, many important bioinformatic tools have been developed for binning these sequencing reads (Morgan, Darling, and Eisen, 2010). Among these, a clad-specific approach implemented in MetaPhlAn is widely used in microbiome studies (Segata et al., 2012).

In addition to what have been described above, each sequencing strategy has pros and cons related to the underlying techniques. Target sequencing is usually cheaper, but only provides information for a certain type of microorganisms. Since only part of the genomes is sequences, as a result, the functionality of the entire community can not be studies using target sequencing data. Clustering of reads into OTU data is another potential caveat. The community profile might change if a different clustering algorithm is applied. Shotgun metagenomic sequencing, on the other hand, provides opportunities to study the function of microbiota by first assembling sequencing reads into individual genomes. Similar to target sequencing approach, it requires an extra step before estimating the abundance of each species. There are multiple computing algorithms to choose from, which can lead to discrepancies in the community profiles or abundances. Shotgun metagenomic sequencing can be used in *de novo* assembly, which often return many metagenome assembled genomes (MAGs), many of these MAGs represent unknown species.

1.3. Current statistical methods in microbiome studies

Due to different sequencing depths and total number of sequence reads in different samples, the read counts are often converted into relative abundance vector with a unit sum. The key property of microbiome data is compositional, meaning that we only obtain relative abundance of each microbe that sums up to 1 for a given sample. Special treatment should be taken before running hypothesis testing or building statistical models with such compositional data. John Atchison, a key statistician

in the field of compositional data analysis, stated that three principles need to be satisfied when performing compsitional data analysis: scale invariance, permutation invariance, and subcompositional coherence (Aitchison, 1982). Bearing these principles in mind, new statistical methods have been developed for microbiome compositional data (Li, 2015).

Another important feature of microbiome compositional data is high dimensionality. Usually, the number of species is more than the number of samples in a study. It poses further challenges in statistical analysis of such data. Recent statistical methodology has focused on the compositional and high-dimensional feature of microbiome data. For example, Rivera-Pinto et al., 2018 proposed to tackle the effect of a subcomposition from the compositional data by best subset selection (Rivera-Pinto et al., 2018). The resulting single measure is defined as balance. Balance is defined as a weighted log ratio between two partitions of a compositional vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$. Let $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ be a p-dimensional indicator vector with three values, 1, -1, and 0, where $z_i = 1$ indicates that the corresponding part in the composition x_i is in the numerator of a ratio, $z_i = -1$ indicates the denominator; and $z_i = 0$ indicates the parts that are not a member of the balance. Let z_+, z_-, and z_0 be non-overlapping sets of indices of the elements in $\mathbf{z}$ whose values are 1, -1 and 0, respectively, and $z_+ \cup z_- \cup z_0 = \{1, 2, \ldots, p\}$. Given $\boldsymbol{z}$, the balance of $\mathbf{x}$ is defined as

$$B_z = \sqrt{\frac{m_+ m_-}{m_+ + m_-}} \log \frac{\left(\prod_{i \in z_+} x_i\right)^{1/m_+}}{\left(\prod_{i \in z_-} x_i\right)^{1/m_-}}, \tag{1.1}$$

where m_+, m_-, m_0 denotes the number of indices in z_+, z_-, z_0 respectively. Such a balance measure provides a scalar summary of the microbial community and can be used to measure community dysbiosis.

The focus of this dissertation is develop a Bayesian framework for balance regression and balance mediation analysis in order to identify the balance indicator vector $\mathbf{z}$ and its corresponding balance B_z that is associated with an outcome or the balance that serves as a mediator in linking treatment to an outcome. Bayesian approaches are chosen due to their ability of exploring the very large space of all possible balance measures that can be constructed from a p dimensional compositonal vector.

1.4. Organization of dissertation

This thesis is organized as follows. In Chapter 2, we develop a Bayesian balance linear regression and an efficient MCMC computational method to identify the balance that is associated with an outcome using stochastic search. In this project, we propose a variable selection algorithm under the Bayesian linear regression framework. Selecting which microbes composing the outcome-associated balance is a NP-hard search problem and we resolve it by using stochastic search. We formulate the search space using an indicator vector with three discrete values that represent the position of corresponding microbes in the balance. The value of indicator vector then determines the balance in each sample. We couple the search procedure with a regression model and design a new proposal distribution for discrete-valued vectors in the stochastic search. To speed up the convergence, we eliminate the Gibbs step by integrating out the regression coefficients from the target distributions. The numerical properties of the proposed method are evaluated by simulations. We apply the balance regression to several real datasets and demonstrate the applicability and findings of the proposed method.

In Chapter 3, we develop a Bayesian balance regression for binary outcome and extend the proposed stochastic search variable selection method for binary balance regression. Instead of using logit function, we propose to use probit function in the regression. There are several advantages of using the probit link function with binary outcome data under the Bayesian framework. In the probit regression, we augment the microbiome data by introducing a latent continuous outcome variable for each sample. The sign of the latent variable determines the actual binary value of the outcome. We then develop a MCMC algorithm by sampling the latent variables along with the search for the balance indicator vector. Numerical studies and real data applications show promising results of the proposed method with binary outcomes.

In Chapter 4, we extend balance regression to balance-based causal mediation analysis where the microbiome composition serves as a mediator. The aim of our proposed method is to find the effect of a treatment on the microbiome composition that also affects the continuous outcome. In traditional mediation analysis, the mediator is usually a single variable. For microbiome studies, the mediator is a compositional vector. The key of our approach is to identify balance in order to reduce the effects from high-dimensional compositional mediator into a single measure of balance. We

propose a generative Bayesian mediation model to estimate the mean vector in the treatment and control groups, as well as the indicator vector for balance. The mediation effect can be calculated by applying the balance indicator vector to the difference between the posterior means of the additive logration transformation of the compositions between treatment and control groups. Inference for the mediation effect can be performed with posterior distributions of the indirect effect.

Finally, in Chapter 5, some future directions to extend the balance analysis to other problems are discussed.

CHAPTER 2

BAYESIAN LINEAR BALANCE REGRESSION

2.1. Introduction

Advances in sequencing technology have enabled researchers to study microbial communities without having to culture them in the laboratory. Recent large scale microbiome studies have greatly expanded our knowledge of the role of microbiome in human health and disease. For example, dysbiosis of the gut microbiome has been linked to Crohn's disease, cancer and cardiometabolic syndrome (Halfvarson et al., 2017; Lewis et al., 2015). Microbial community composition from a sample can be inferred from16S rRNA sequencing, which provides data on read counts that are assigned to different taxa. However, the total count is not informative because of the sampling, sample preparation and sequencing processes. These counts are used to obtain estimates of the relative abundance of the p taxa in the community, which is compositional with a unit sum constraint. Such data are called compositional data with p parts, which can be summarized as $\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{S}^{p-1}$, where $x_i \geq 0, i = 1, 2, \cdots, p$, $\sum_{i=1}^{p} x_i = 1$ and $\mathbb{S}^{p-1}$ represents $p-1$-dimensional simplex.

Standard statistical methods cannot be applied directly to microbiome compositional data because of the dependence among the p parts (Li, 2015). In addition, according to (Aitchison, 1982), the analysis of compositional data should satisfy two principles: *scale invariance and subcompositional coherence*. The former principle implies that the total count or sequencing read count in microbiome study should be irrelevant in statistical analysis of the microbial composition. Subcompositional coherence requires that even two microbial communities have different numbers of bacteria, any associations found among the overlapping taxa should be not affected by other non-overlapping taxa. This principle is important, since in microbiome studies, researchers rarely identify the same number of taxa in different studies.

Log-ratio analysis aims to preserve scale invariance and subcompositional coherence in compositional data analysis. Such a transformation is often performed by first choosing a component from $\mathbf{x}$ as the denominator, say x_1. Log ratios are then computed with one of the remaining components

as the numerator,

$$\log \frac{x_i}{x_1} \qquad \forall i = 2, 3, \ldots, p.$$

However, the choice of the denominator is arbitrary, so this log-ratio transformation is not symmetric and may yield different results for different choices of the denominator when these log ratios are used as covariates in regression analysis. Recently, several researchers Lu, Shi, and Li, 2019; Shi, Zhang, and Li, 2016; Wang and Zhao, 2017 have developed new inference approaches that alleviate the non-symmetric issue by introducing a constraint on the regression coefficients. Aitchison Aitchson, 2003 also proposed a symmetric log-ratio transformation, called centered log-ratio transformation and is defined as

$$\log \frac{x_i}{g(\mathbf{x})} \qquad \forall i = 1, 2, \ldots, p,$$

where $g(\mathbf{x})$ is the geometric mean of all p components. Centered log-ratio transformed compositional data are intrinsically colinear and the resulting matrix is singular, which can lead to problems in downstream data analysis. Bates and Tibshirani Bates and Tibshirani, 2019 proposed a penalized variable selection procedure for log ratios formed by all pairwise parts in a compositional data. The total number of log-ratios is however very large.

All the above log-ratio transformations are developed to address the foundation of compositional data analysis that individual components have no meaning and it is their relationship to other components that has meaning. In microbiome studies, the community members may exhibit co-operative or co-exclusive relationship and the entire community maintains a homeostasis (Griffin, West, and Buckling, 2004; West et al., 2006). Disruptions in relative proportions of certain microbes can cause perturbed homeostasis and even pathogenesis. For example, the ratio between *Firmicutes* and *Bacteroidetes* is significantly higher in overweight and obese subjects (Castaner et al., 2018). Thus, it is critical to carry out statistical analysis based on ratios between groups of taxa, which is more informative than only considering a pair of taxa and is capable of capturing the dysbiosis of a microbial community in pathogenic status.

Log-ratio of partitions of a composition, termed balance, was first introduced in geology by Egozcue and Pawlowsky-Glahn (Egozcue and Pawlowsky-Glahn, 2005), and was recently applied to microbiome studies with meaningful discoveries (Morton et al., 2017). Balance is defined as a weighted

log ratio between two partitions of a compositional vector. Let $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ be a p-dimensional indicator vector with three values, 1, -1, and 0, where $z_i = 1$ indicates that the corresponding part in the composition x_i is in the numerator of a ratio, $z_i = -1$ indicates the denominator; and $z_i = 0$ indicates the parts that are not a member of the balance. Let z_+, z_-, and z_0 be non-overlapping sets of indices of the elements in $\mathbf{z}$ whose values are 1, -1 and 0, respectively, and $z_+ \cup z_- \cup z_0 = \{1, 2, \ldots, p\}$. Given $\boldsymbol{z}$, the balance is defined as

$$B_z = \sqrt{\frac{m_+ m_-}{m_+ + m_-}} \log \frac{\left(\prod_{i \in z_+} x_i\right)^{1/m_+}}{\left(\prod_{i \in z_-} x_i\right)^{1/m_-}}, \tag{2.1}$$

where m_+, m_-, m_0 denotes the number of indices in z_+, z_-, z_0 respectively. In previous applications, the value of the indicator vector $\mathbf{z}$ is chosen *a priori* based on a phylogenetic tree or some other biological knowledge (Morton et al., 2017; Washburne et al., 2017). However, such information may be biased or even misleading since the taxonomic positions do not always differentiate beneficial from detrimental taxa.

One important question in microbiome data analysis is to identify the vector $\mathbf{z}$ and the corresponding balance that is associated with a certain outcome variable based on the data collected. Rivera-Pinto et al. Rivera-Pinto et al., 2018 proposed a forward selection procedure to derive a balance from microbiome data. Their method adds one 'best' taxon at a time into the balance based on an optimization criterium such as the mean square error and stops when a prescribed maximum number of taxa is achieved or the improvement of the optimization parameter is lower than a threshold. However, such a search only explores a small subspace of all possible balances and can lead to sub-optimal results if the optimal model is not visited in the search space. In addition, when p is large, searching for all possible 3^p balance becomes infeasible.

In this paper we propose a Bayesian balance-regression to construct a balance and to model the association between the balance and outcome of interest using stochastic search. Unlike the search method proposed by Rivera-Pinto et al., 2018, we develop a Markov chain Monte Carlo (MCMC) stochastic search algorithm that is capable to explore the large search space. The final structure of the balance can be inferred from the marginal posterior probability of each z_i or from the most probable configuration of $\mathbf{z}$ when the MCMC algorithm converges. The MCMC algorithm enables the posterior inference for the regression coefficients conditional on the structure of balance at each

iteration of the search algorithm. Since the latent vector $\mathbf{z}$ and therefore the balance is unknown, our MCMC algorithm is very different from the standard Bayesian regression (George and McCulloch, 1993) where the covariates are known. In our case, besides sampling the model parameters, we have to sample the latent indicator vector $\mathbf{z}$. In Bayesian variable selection, the covariates are observed and one samples the covariate-inclusion indicators to select the covariate that are associated with the response.

In this Chapter we propose a Bayesian treatment, based on the Bayesian regression framework (George and McCulloch, 1993), to construct a balance and model the association between balance and outcome of interest using stochastic search. Besides specifications for priors of unknowns and setting a threshold for marginal posterior inclusion probability for each taxon, our procedure is fully automatic. Unlike the greedy search proposed by Rivera-Pinto et al., 2018, we design a proposal that is capable to explore the entire search space and achieve global optimum. Structure of the balance can be inferred from the marginal posterior inclusion probability or be reported from the most probable configurations when the search algorithm converges. We propose to perform posterior inference for the regression coefficients conditional on the structure of balance at each iteration of the search algorithm. As a result, the inference on regression coefficients are averaged over the most probable structures of a balance.

2.2. Bayesian linear balance regression

Consider a microbiome study with n $i.i.d.$ samples from a population. Let $\mathbf{y} = (y_1, \cdots, y_n)$ be the vector of a continuous outcome of length n, $\mathbf{X} = \{x_{ij}, i = 1, \cdots, n, j = 1 \cdots, p\}$ be the compositional matrix with dimension $n \times p$, where p is the number of taxa observed in the samples and x_{ij} is the abundance of the jth taxon in the ith sample estimated from the sequencing counts. Let $\mathbf{z}$ represent the indicator vector that determines the structure of the balance and the resulting balance B_z as defined in equation (2.1). We can calculate the balance using (2.1) for the entire sample when $\mathbf{z}$ is known. For a given balance configuration $\mathbf{z}$, let $\mathbf{B}_z$ be the vector of the balance values for the n individuals.

Our main interest is to identify a $\mathbf{z}$ indicator vector and the corresponding balance that is strongly associated with the outcome. Let $\mathbf{B_{1z}} = (\mathbf{1}, \mathbf{B}_z)$ be the $n \times 2$ matrix of columns of $\mathbf{1}$ and the vector

$\mathbf{B}_z$. For a given balance configuration $\boldsymbol{z}$, we define the following Bayesian balance-regression

$$\mathbf{y}|\mathbf{z} = \mathbf{B_{1z}}\boldsymbol{\beta_z} + \boldsymbol{\epsilon_z}, \qquad \boldsymbol{\epsilon_z} \sim MVN(0, \sigma^2\mathbf{I}) \tag{2.2}$$

where $\boldsymbol{\beta}_z = (\beta_{0z}, \beta_{1z})^T$ is the regression coefficients that depend on $\boldsymbol{z}$. To complete the model, we assume the following prior distributions for all the unknown parameters, namely $\boldsymbol{\beta_z}, \sigma^2, \mathbf{z}$, in order to perform a posterior inference. Similar to Bayesian regression, we assume a normal-inverse-gamma distribution for $\boldsymbol{\beta}_z$ and σ^2 as discussed in George and McCulloch, 1993. More specifically, we assume that $\boldsymbol{\beta}_z = (\beta_{0z}, \beta_{1z})^T$ are jointly normal with mean $\mathbf{b_0} = (b_{00}, b_{01})^T$ and covariance $\sigma^2\mathbf{V}$, where $\mathbf{V}$ is a diagonal matrix with entries $h,$ and c. We also assume that σ^2 follows an inverse Gamma distribution with parameters $(v/2, v\lambda/2)$. The advantage to use a normal-inverse-gamma prior is that we can obtain a closed form of the marginal distribution $\mathbf{y}|\mathbf{z}$ after integrating out $\beta_{\mathbf{z}}$ and σ^2. We can therefore sample $\mathbf{z}$ with shorter chains and faster convergence, as compared with other choices of the prior distributions.

The balance configuration vector of $\mathbf{z}$ determines the search space and the prior on $\mathbf{z}$ represents our belief of the true structure of a balance. A straightforward choice is an independent multinomial distribution for each component of $\mathbf{z}$,

$$f(\mathbf{z}) = w_1^{m_+} w_2^{m_-} (1 - w_1 - w_2)^{p - m_+ - m_-}, \tag{2.3}$$

where w_1 and w_2 are the expected number of variables in z_+ and z_- group respectively. They affect the sizes of taxa with positive and negative effects in the posterior inference of the balance. A non-informative choice is $w_1 = w_2 = 1/3$, resulting in $f(\mathbf{z}) = (1/3)^p$. On the other hand, small values of w_1 and w_2 assume sparse structure of the balance. This prior has implicit assumptions that the sizes of the sets z_+ and z_- are approximately equal. It is possible to relax the independence specification and use a dependent prior among elements of $\mathbf{z}$. If p is large, we suggest to use the sparse prior for the balance configuration $\boldsymbol{z}$.

2.3. Posterior inference based on MCMC

Based on the Bayesian balance-regression model and the respective prior distributions of the parameters, the posterior distribution $\mathbf{z}|\mathbf{y}$ is proportional to the joint distribution of $(\mathbf{y}, \mathbf{z})$:

$$f(\mathbf{z}|\mathbf{y}) \propto |\mathbf{V}^*|^{1/2} \left(v\lambda + \mathbf{b_0}^T \mathbf{V}^{-1} \mathbf{b_0} + \mathbf{y}^T \mathbf{y} - \mathbf{u}^T \mathbf{V}^{*-1} \mathbf{u} \right)^{-(v+n)/2} f(\mathbf{z}) \tag{2.4}$$

where $\mathbf{V}^* = (\mathbf{V}^{-1} + \mathbf{B_{1z}}^T \mathbf{B_{1z}})^{-1}$ and $\mathbf{u} = \mathbf{V}^*(\mathbf{V^{-1}b_0} + \mathbf{B_{1z}^T y})$ and $\boldsymbol{\beta}_z$ is integrated out.

We develop the following MCMC algorithm to explore high dimensional sample space of $\boldsymbol{z}$ and to sample $\mathbf{z}$ from its posterior distribution given in (2.4). For stochastic search, three possible value changes can occur during the random exploration: i) between 0 and 1; ii) between 0 and -1; iii) between 1 and -1. Modifications between two values has three alterations: value conversions in either direction and value switch. We set equal probabilities among all possible modifications, but such probabilities can be adjusted to speed up the exploration of the sampling space. The convergence and mixing of the MCMC algorithm can be evaluated using different chains witrh different starting points and by examining the cumulative moving average for each component of $\mathbf{z}$. After convergence, the posterior inclusion probabilities for the sets z_+, z_-, z_0 are simply taken as the marginal proportions of $1, -1, 0$ in $\mathbf{z}$. A threshold can be set to make inference on which set of particular taxa should belong to the balance and the resulting balance definition.

Conditioning on $\boldsymbol{z}$, we can obtain the predictive posterior distribution of the regression coefficients $\boldsymbol{\beta}_z$. From model (2.2) and the normal prior distribution of $\boldsymbol{\beta}_z$, one can show that $\boldsymbol{\beta}_z|\mathbf{y}, \mathbf{z}$ has a multivariate t-density with degrees of freedom $\lambda + n$, location parameter $\mathbf{u}$, and shape matrix

$$\tilde{\boldsymbol{V}} = \frac{v\lambda + \mathbf{b_0}^T \mathbf{V}^{-1} \mathbf{b_0} + \mathbf{y}^T \mathbf{y} - \mathbf{u}^T \mathbf{V}^{*-1} \mathbf{u}}{v + n} \mathbf{V}^*. \tag{2.5}$$

During each iteration of MCMC, we sample $\boldsymbol{\beta}_z$ from its posterior distribution and perform posterior inference for $\boldsymbol{\beta}_z$ based on these sampling points. Finally, at the convergence, due to randomness of $\boldsymbol{z}$, the posterior samples of $\boldsymbol{\beta}_z$ are averaged over those $\mathbf{z}$, which gives the mean of the predictive posterior distribution of $\boldsymbol{\beta}_z$ based on model average. Algorithm (1) presents an outline of the final MCMC agorithm for fitting the balance-regression.

Algorithm 1: MCMC algorithm for the Bayesian balance-regression analysis

1 sample $\mathbf{z}|\mathbf{y}^*$ based on its posterior probability (2.4) and stochastic search.
2 sample $\boldsymbol{\beta}_z|\mathbf{y}, \mathbf{z}$ from a t-distribution with location parameter $\mathbf{u}$ and shape matrix given in (2.5).
3 Obtain posterior probability of $Pr(z_i|\mathbf{y}, \mathbf{X})$ for balance inference and predictive posterior distribution of $\boldsymbol{\beta}_z$.

Since the target distribution of $\mathbf{z}$ has two modes when the prior of β_{1z} has a mean value of 0. To make it identifiable, if the estimate of β_{1z} is less than 0, we flip the signs of $\mathbf{z}$ and return a positive estimate of β_{1z}. This effectively constrains β_{1z} to be positive in order to make $\mathbf{z}$ identifiable.

2.4. Numerical studies

2.4.1. Data generation

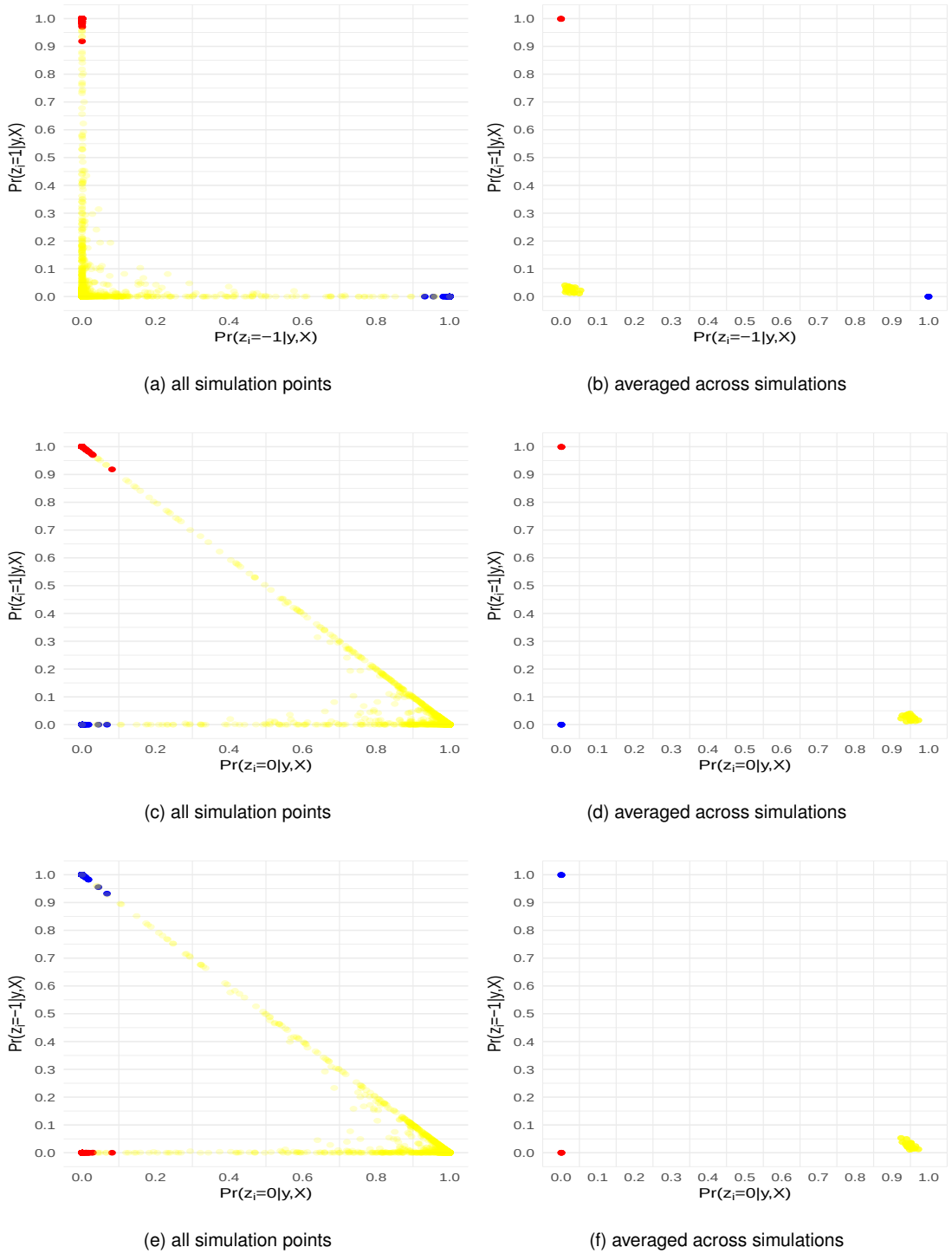
We evaluate the performance of our proposed method using simulations, For each simulation, we set $n = 100, \quad p = 30, \quad \beta_0 = 1, \quad \sigma^2 = 1$ and β_1 taking three values 1, 0.5 and 0.1 representing strong, moderate and small effect of the balance on the outcome. The hyperparameters for prior distribution of $\boldsymbol{\beta}$ are chosen as $b_{00} = 0, \quad b_{01} = 0, \quad c = 10^6, \quad h = 10^6$, representing uninformative prior information. For binary outcome, we set $c = 10^3, \quad h = 10^3$ due to the restrictions in Gibbs sampling of the latent vector $\boldsymbol{y}^*$ and set w_1 and w_2 to be $1/3$, representing each component being equally likely to be in the z_+, z_-, z_0 sets. The Inverse Gamma prior distribution $IG(v/2, \lambda/2)$ corresponds to a likelihood of σ^2 that comes from v independent $N(0, \lambda)$ observations. As a result, v can be treated as the prior sample size and λ as a prior estimate for σ^2. In practice λ can be chosen as the sample variance of the outcome variable in model (2.2).

To generate the compositions, we first generate n independent vector of dimension p from a multivariate normal with zero mean, variance of 10 and equicorrelation of 0.2. Taxon count matrix is generated by exponentiating the above matrix and then taking the greatest integer smaller or equal to the respective numbers. Zero counts correspond to rare taxa and are replaced with 0.5. Finally the count matrix is normalized per row and to obtain the compositional matrix $\mathbf{X}$. Across all simulations, the first 3 taxa are in the z_+ set of a balance and the next 3 taxa are in z_- set. Outcomes are generated according to model (2.2). The initial values for $\mathbf{z}$ are chosen randomly, The performance of mixing of the Markov chains is evaluated with several starting points.

2.4.2. Simulation results under uniform prior

Figures 2.1, 2.2,2.3,2.4 summarize the posterior probabilities $Pr(z_i = 1|\mathbf{y}, \mathbf{X})$, $Pr(z_i = -1|\mathbf{y}, \mathbf{X})$ and $Pr(z_i = 0|\mathbf{y}, \mathbf{X})$ after 10^5 iterations after 10^4 burn-ins for all 30 taxa across 100 simulations under the uniform prior assumption. When the balance effect is strong, the posterior inclusion probabilities for the first 3 taxa are above 0.75 across all simulations for the compositional taxa in the z_+ set and near 0 for those in the z_- group. Similarly, compositional taxa 4 to 6 can be correctly identified as in the z_- set with high probabilities.

When the balance effect becomes weaker, the elements in those two sets can sometimes be identified with high probabilities. However, the posterior inclusion probabilities are estimated toward the prior due to a low signal-to-noise ratio as compare to those models with $\beta_{1z} = 1$. The remaining taxa that are not part of the true balance index can sometimes be estimated to have higher posterior probability of being in z_+ or z_- set.



(a) all simulation points

(b) averaged across simulations

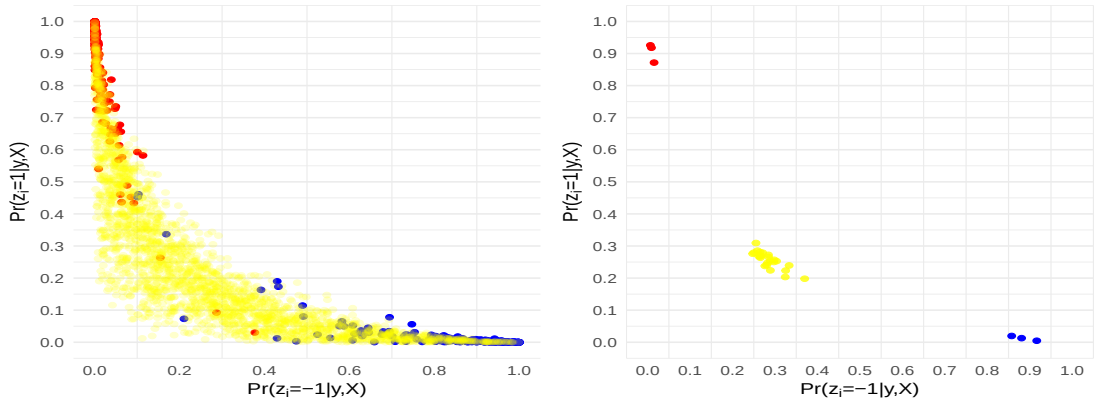
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

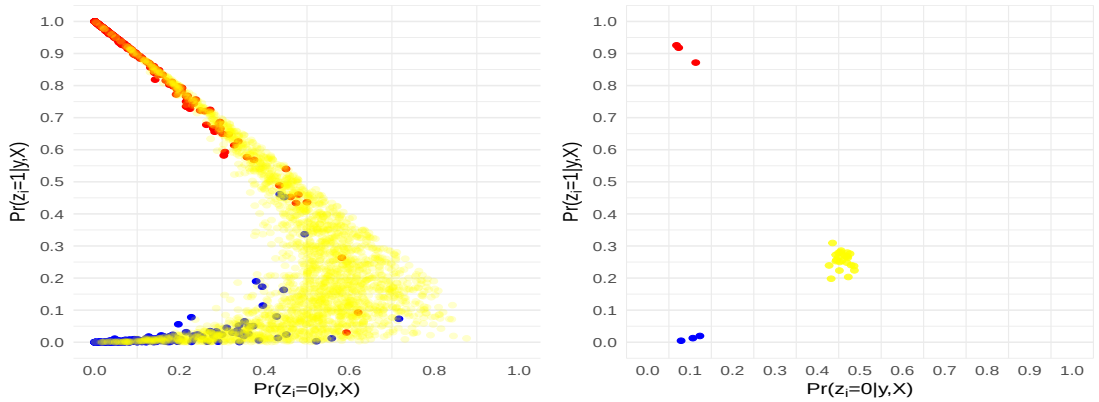
(f) averaged across simulations

Figure 2.1: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 1$ under uniform prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



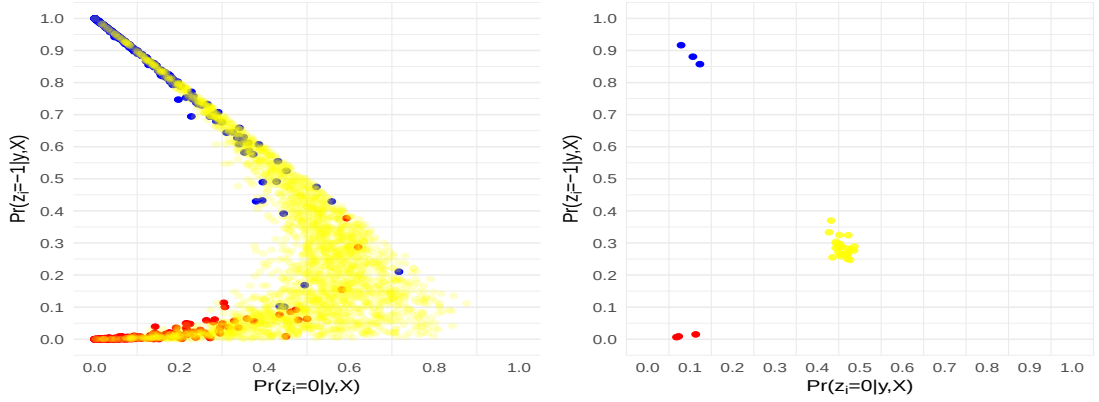
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

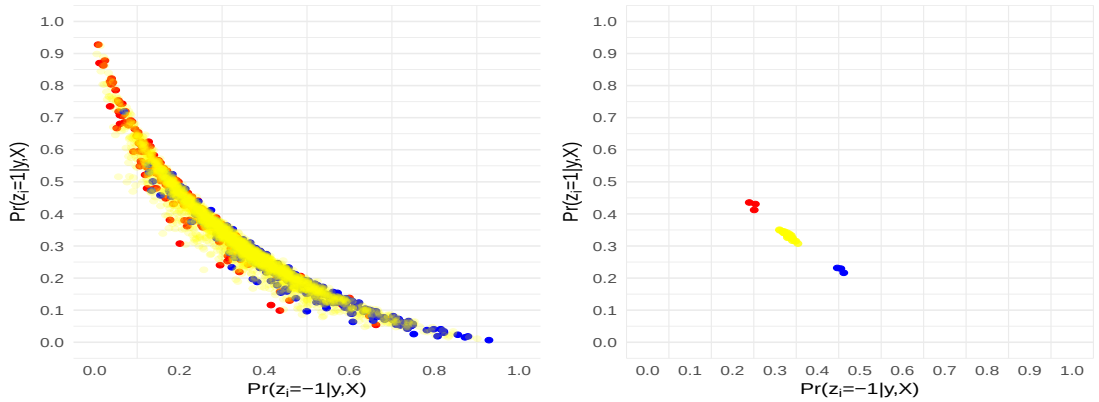
(d) averaged across simulations



(e) all simulation points

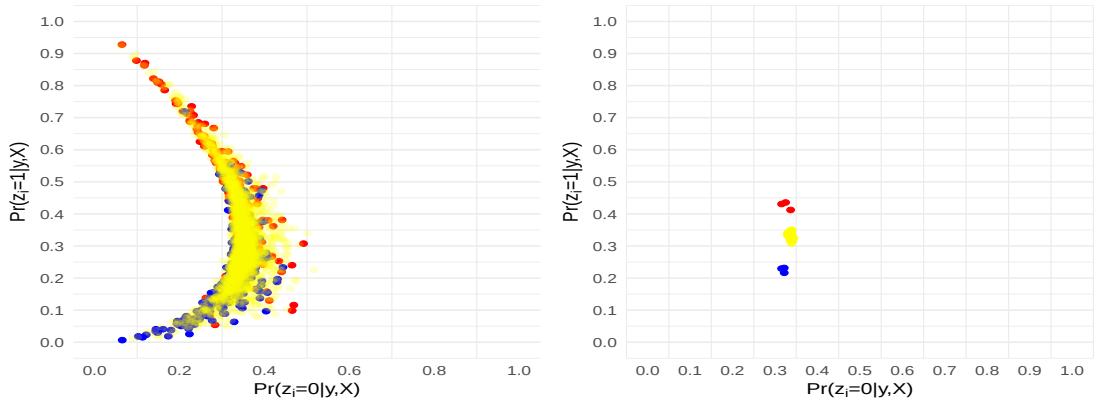
(f) averaged across simulations

Figure 2.2: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0.5$ under uniform prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



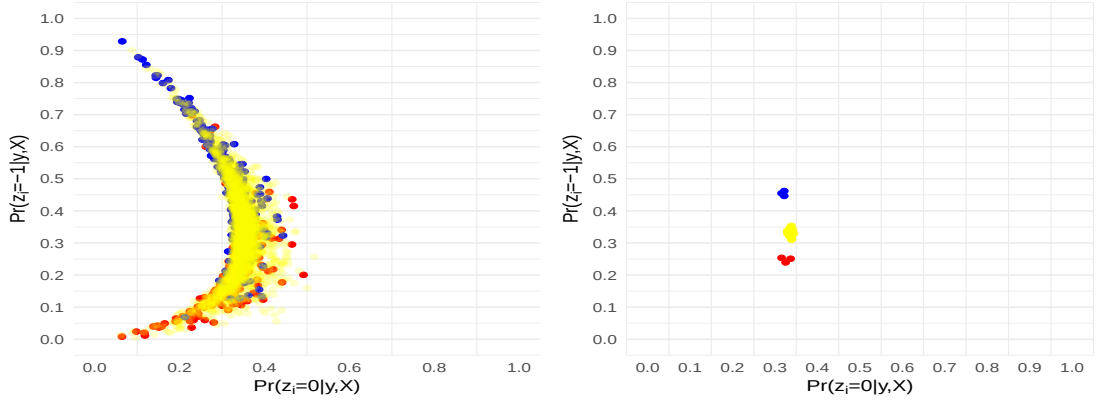
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

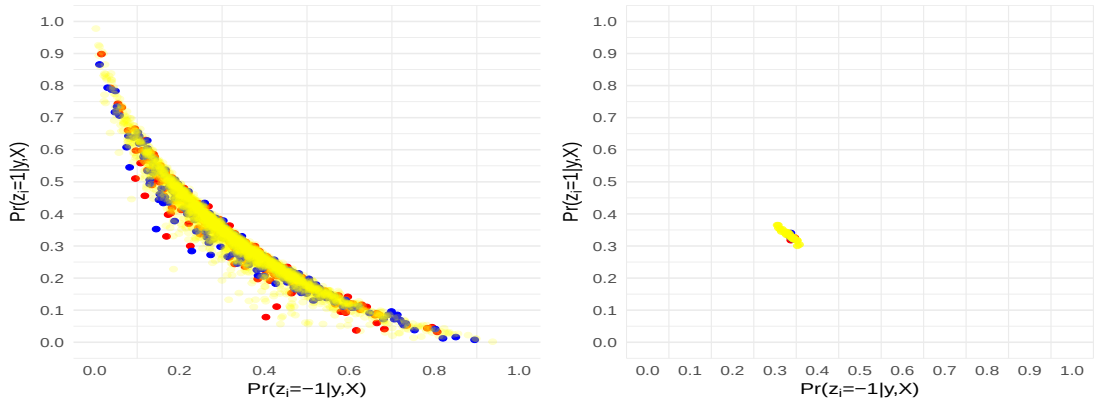
(d) averaged across simulations



(e) all simulation points

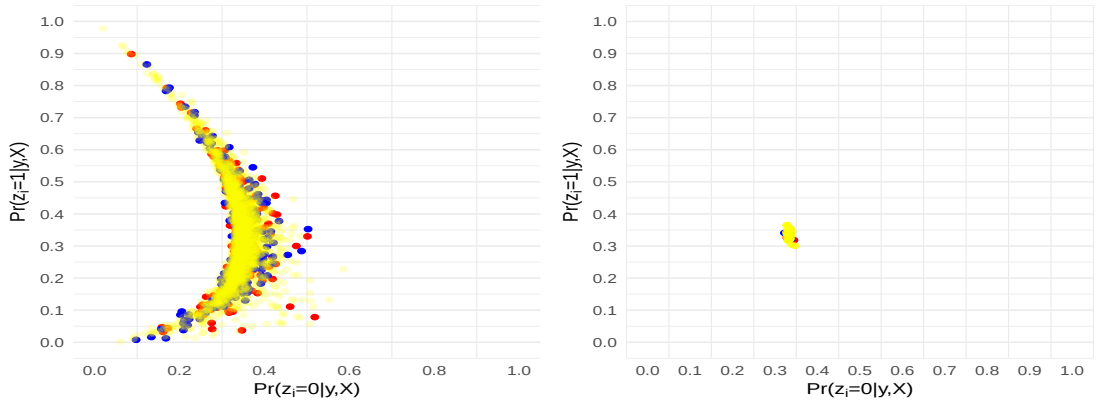
(f) averaged across simulations

Figure 2.3: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0.1$ under uniform prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



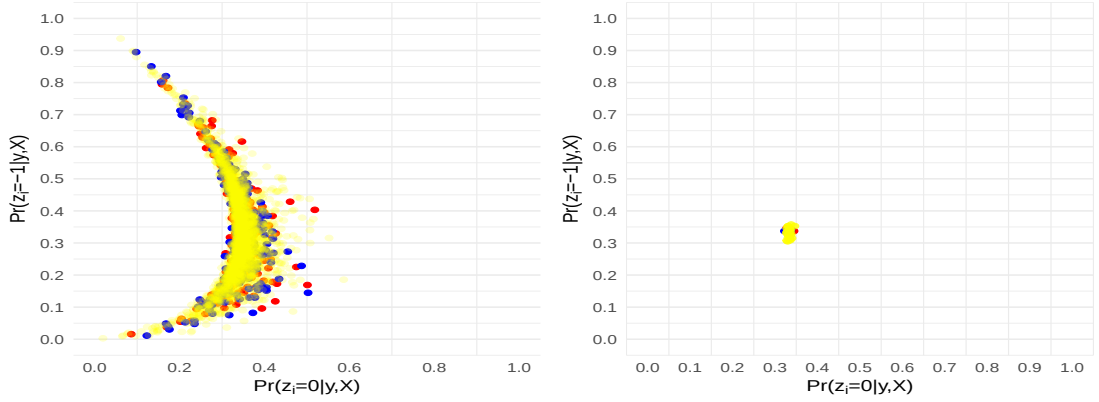
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

(d) averaged across simulations



(e) all simulation points

(f) averaged across simulations

Figure 2.4: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0$ under uniform prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).

Figure 2.5 shows the posterior mean of β_{1z} over 100 simulations. These boxplots show that they are centered around the true values. All of the 95% empirical credible intervals derived from the sampling points contain the true value of β_{1z}, indicating that the MCMC algorithm provides valid estimate of the balance effect.

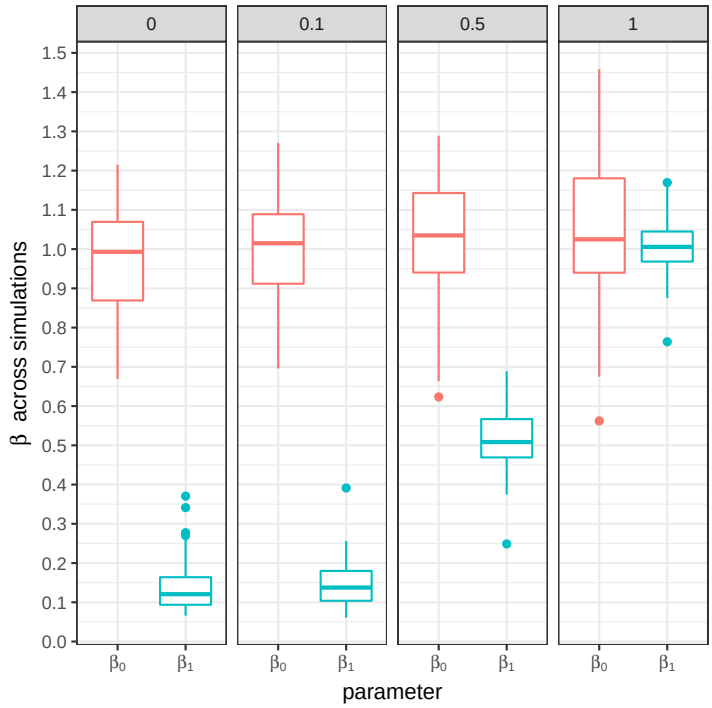
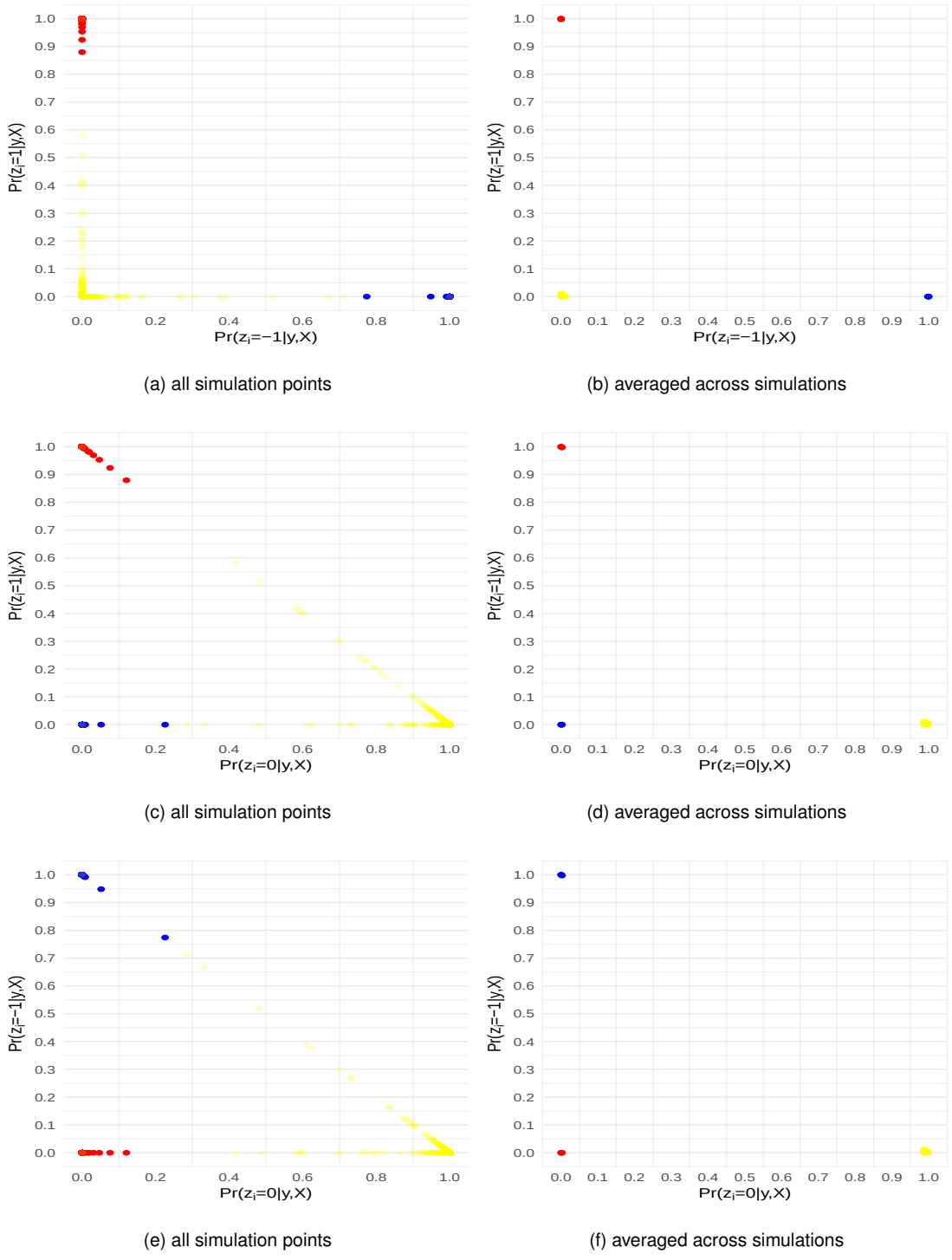


Figure 2.5: Posterior mean of β_{1z} over 100 simulations in balance linear regression under uniform prior.

2.4.3. Simulation results under sparse prior

Figures 2.6 ,2.7,2.8 and 2.9 summarize the posterior inclusion probabilities after 10^5 iterations after 10^4 burn-ins for all 30 taxa across 100 simulations under the assumption of sparse prior. Similar to the uniform prior, when the effect size is strong $\beta_{1z} = 1$ and medium $\beta_{1z} = 0.5$, all taxa can be correctly identified in the balance. When the effect size is small $\beta_{1z} = 0.1$ and null $\beta_{1z} = 0$, averaged posterior probabilities for each taxon are estimated toward the prior and almost all of them are identified in the z_0 set. Compared to uniform prior, all the simulation points are more dispersed.



(a) all simulation points

(b) averaged across simulations

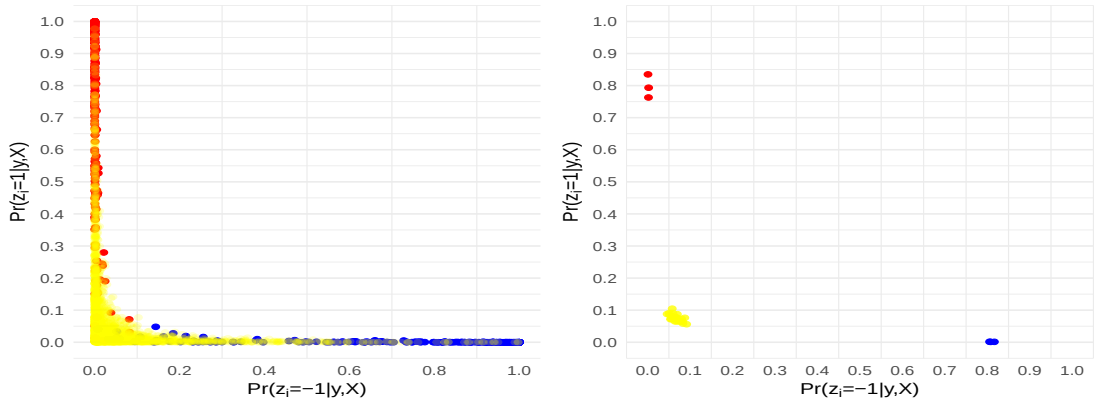
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

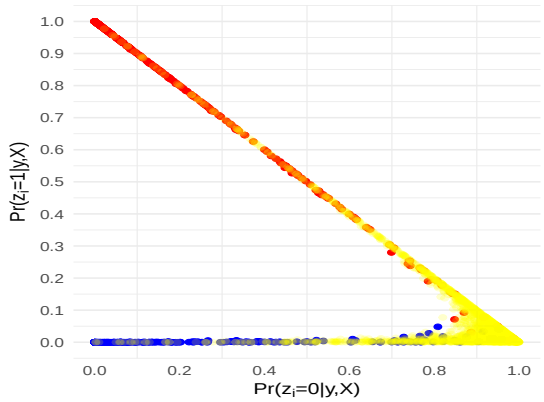
(f) averaged across simulations

Figure 2.6: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 1$ under sparse prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).

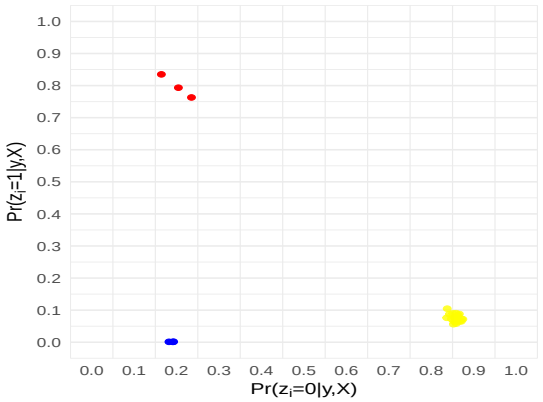


(a) all simulation points

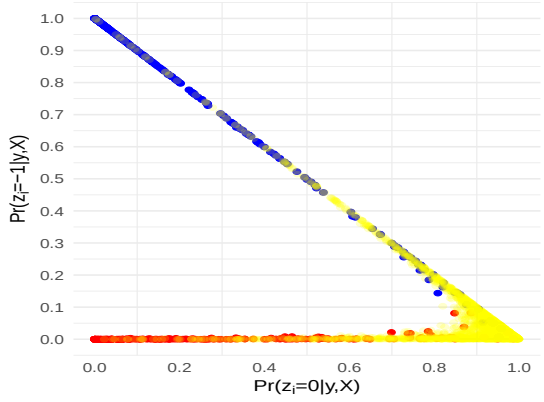
(b) averaged across simulations



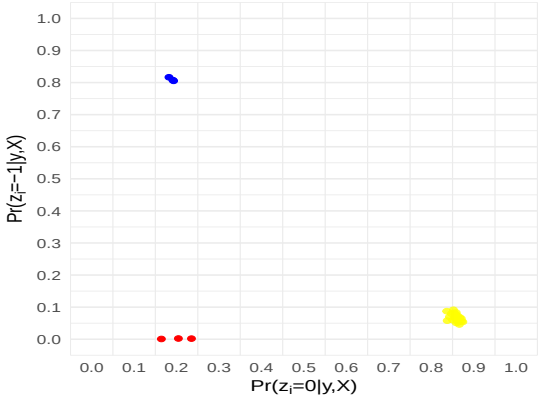
(c) all simulation points



(d) averaged across simulations

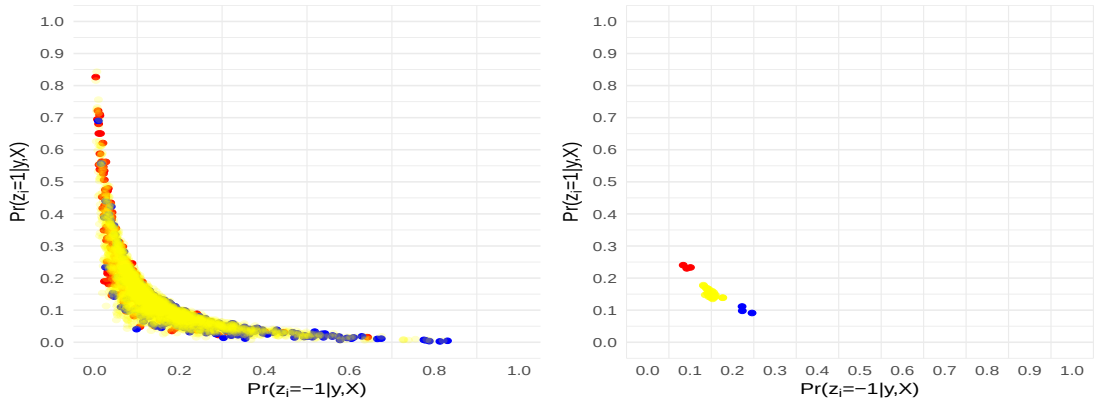


(e) all simulation points



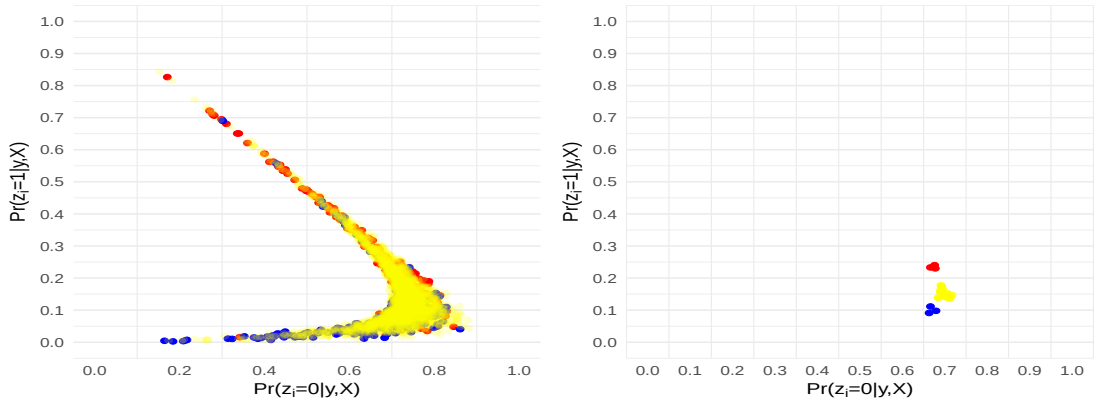
(f) averaged across simulations

Figure 2.7: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0.5$ under sparse prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



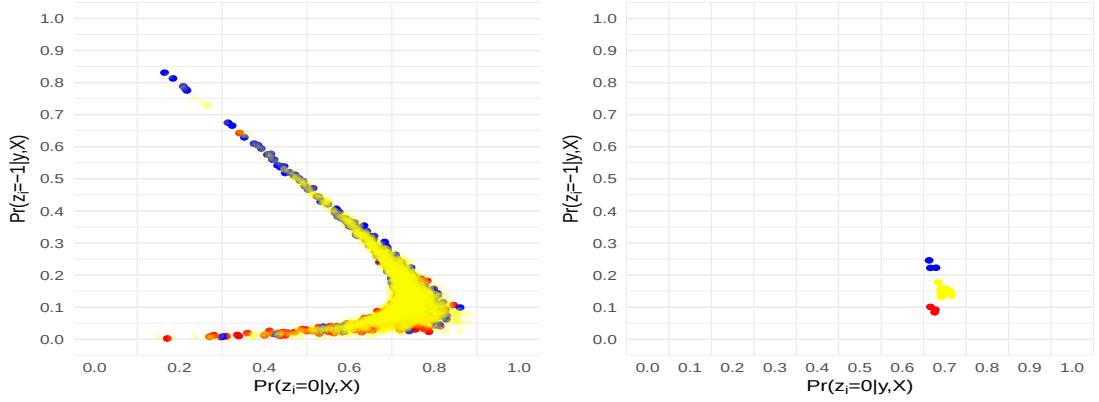
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

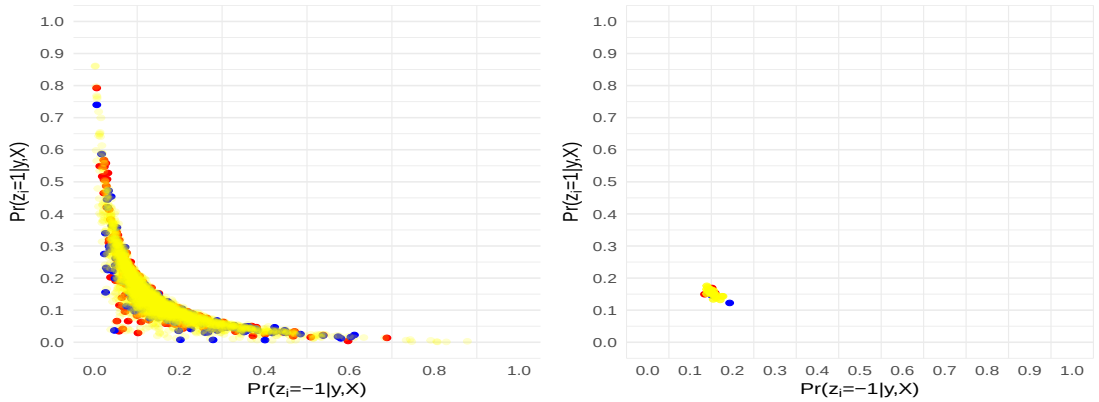
(d) averaged across simulations



(e) all simulation points

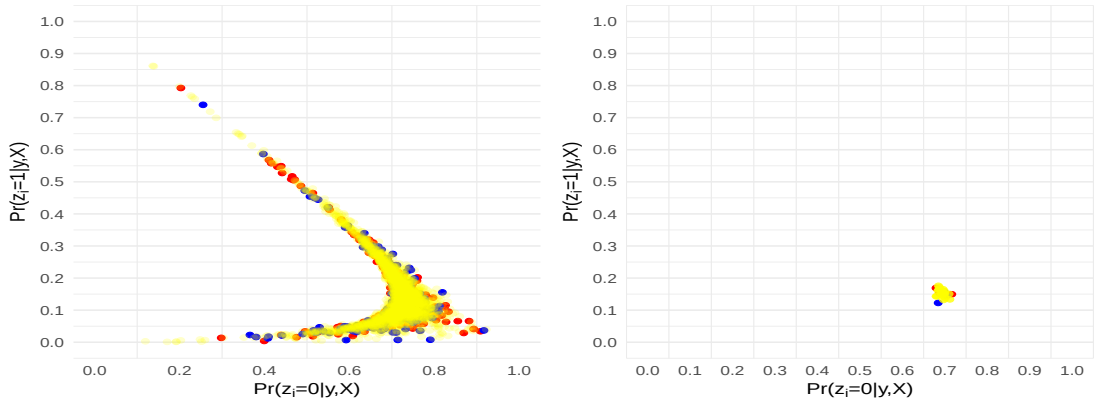
(f) averaged across simulations

Figure 2.8: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0.1$ under sparse prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



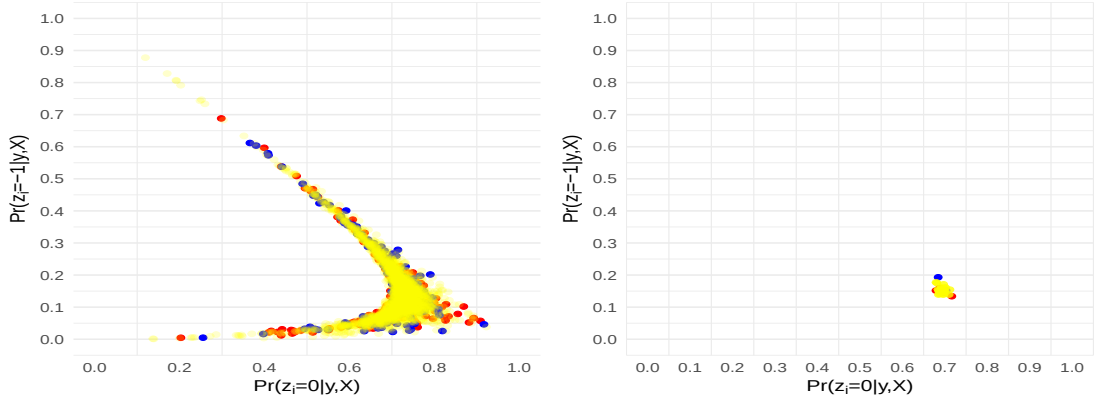
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

(d) averaged across simulations



(e) all simulation points

(f) averaged across simulations

Figure 2.9: Simulation results for balance linear regression model with balance effect $\beta_{1z} = 0$ under sparse prior. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).

Figure 2.10 shows the posterior mean of β_{1z} over 100 simulations. These boxplots show that they are centered around the true values. All of the 95% empirical credible intervals derived from the sampling points contain the true value of β_{1z}, indicating that the MCMC algorithm provides valid estimate of the balance effect.

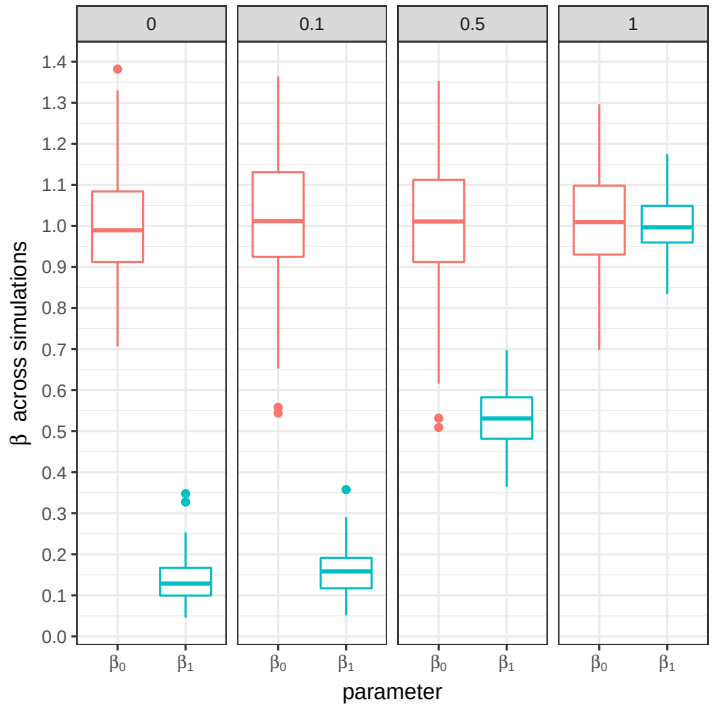


Figure 2.10: Posterior mean of β_{1z} over 100 simulations in balance linear regression under sparse prior.

2.5. Application to two real studies

2.5.1. Association between gut microbiome and BMI - analysis of COMBO Data

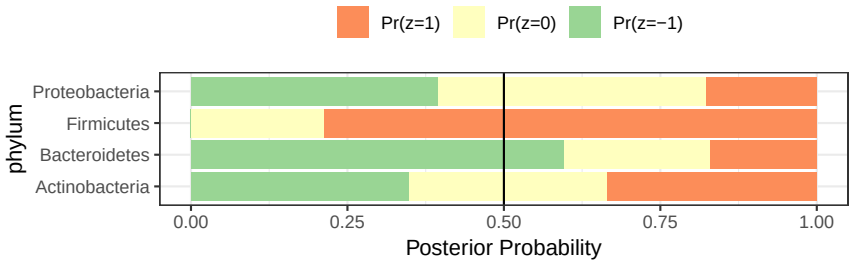
Wu et al., 2011 conducted a $16S$ target sequencing of gut microbiota, aiming to associate the microbial compositional profiles to diet. In this cross-sectional study, 98 healthy volunteers were recruited and their fecal samples were collected. Sample DNA were amplified according to the V1-V2 region in the $16S$ ribosomal DNA. The operational taxonomic units (OTUs) were grouped into 87 taxa, which had at least 0.2% relative abundance in one sample and appeared in more than 10% samples. It has been reported previously that microbiome compositions were associated with obesity (Shi, Zhang, and Li, 2016). We refer this dataset as 'COMBO' data. In this analysis, we aim

to find the balance that is associated with the outcome BMI.

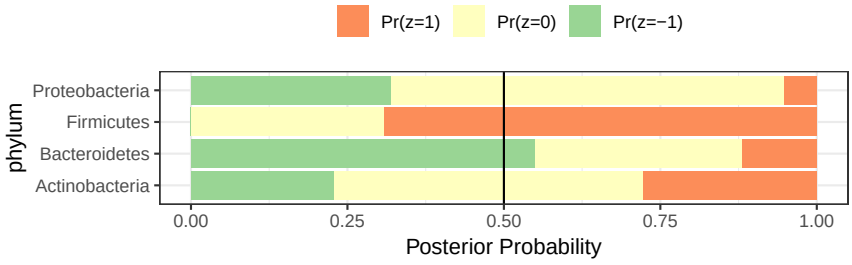
We aggregate the count data into two different taxonomic levels: genus level and phylum level. Relative abundance is obtained by normalizing the counts. A total of 45 genera and 4 phyla are identified in the data. The proposed MCMC approach is carried out with the same values of hyperparameters as in Section 2.4. We run 5 chains with different starting values to assess the convergence. Burn-in step is taken to be 10^4 and total iteration is taken to be 5×10^4. The total number of iterations is more than necessary for the phylum level data. In practice, 10^3 is sufficient for a 4-dimensional $\boldsymbol{z}$ to converge to the posterior distribution. We perform the analysis with both sparse and uniform search space prior where $w_1 = w_2 = 0.1$ and $w_1 = w_2 = 1/3$ respectively, to test the sensitivity of posterior inference from different prior choices.

Analyses of COMBO data at phylum level

Figure 2.11 and Table 2.3 show the analysis results using the phylum level data. A total of 4 phyla are included in COMBO data so the balance indicator $\boldsymbol{z}$ has a dimension of 4. There are a total of 81 possible values of $\boldsymbol{z}$ and the Markov chain quickly converged. Figure 2.11 shows the posterior probability for each taxa being in z_+, z_-, z_0 sets. To determine which taxa should be included in balance, maximum a posteriori (MAP) is used to infer which set of z_+, z_-, z_0 that each taxon belongs to. In our analysis, we use 0.5 as the threshold to determine which set each taxa belongs to. Compared to MAP, using 0.5 as a threshold results in fewer taxa in the calculation of balance index. The reason of using 0.5 threshold instead of MAP is that we require a stronger evidence from the data to infer the balance index. For phylum level COMBO data, the two criteria lead to the same inference regarding balance with either the uniform or the sparse prior (Figure 2.11). The balance is the logratio between Firmicutes and Bacteroidetes. This balance has already been reported in literature to be associated with diabetee and obesity and the underlying mechanism has been elucidated through *in vivo* studies (Backhed et al., 2004; Castaner et al., 2018; Ley et al., 2005b; Turnbaugh et al., 2009; Zhang et al., 2009).



(a) Uniform Prior



(b) Sparse Prior

Figure 2.11: Posterior probabilities for each phylum being in each of the z_+, z_-, z_0 sets under uniform prior (a) and sparse prior (b) for z. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.

Table 2.1 and Table 2.2 present all the posterior probabilities for phylum level COMBO data analysis with uniform and sparse prior using 5 different starting points. The posterior probabilities are fairly consistent using different prior parameters, indicating that our methods are not too sensitive to choices of the initial values.

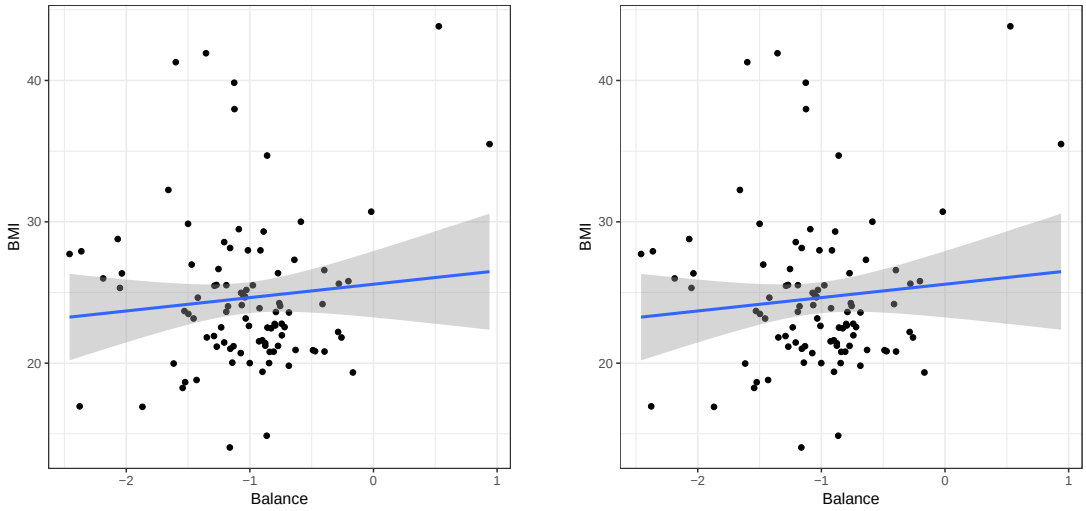
Table 2.1: Posterior probabilities with 5 different starting points. Results are for phylum level COMBO data with uniform prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Actinobacteria	0.34	0.35	0.31	0.33	0.35	0.31	0.33	0.35	0.31	0.34	0.35	0.31	0.34	0.35	0.31
Bacteroidetes	0.17	0.59	0.24	0.17	0.60	0.23	0.17	0.59	0.23	0.17	0.60	0.24	0.17	0.59	0.24
Firmicutes	0.78	0.00	0.22	0.78	0.00	0.22	0.79	0.00	0.21	0.79	0.00	0.21	0.79	0.00	0.21
Proteobacteria	0.18	0.40	0.42	0.18	0.39	0.43	0.18	0.39	0.43	0.17	0.39	0.43	0.18	0.40	0.42

Table 2.2: Posterior probabilities with 5 different starting points. Results are for the phylum level COMBO data with uniform prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Actinobacteria	0.28	0.23	0.50	0.28	0.23	0.50	0.28	0.23	0.50	0.28	0.23	0.50	0.28	0.23	0.49
Bacteroidetes	0.12	0.55	0.33	0.12	0.55	0.33	0.12	0.55	0.33	0.12	0.55	0.33	0.12	0.55	0.33
Firmicutes	0.69	0.00	0.31	0.69	0.00	0.31	0.69	0.00	0.31	0.69	0.00	0.31	0.69	0.00	0.31
Proteobacteria	0.05	0.32	0.63	0.05	0.32	0.63	0.05	0.32	0.63	0.05	0.32	0.63	0.05	0.32	0.63

The scatterplot between BMI and the estimated balance is shown in Figure 2.12, indicating some association between the balance identified and BMI. The posterior distributions for the regression coefficients are summarized in Table 2.3, showing similar results using two different priors. This indicates the signal in the phylum level data is very strong.



(a) Uniform Prior

(b) Sparse Prior

Figure 2.12: BMI vs the estimated balance in COMBO data. Blue line represents the fitted line from the linear least squares.

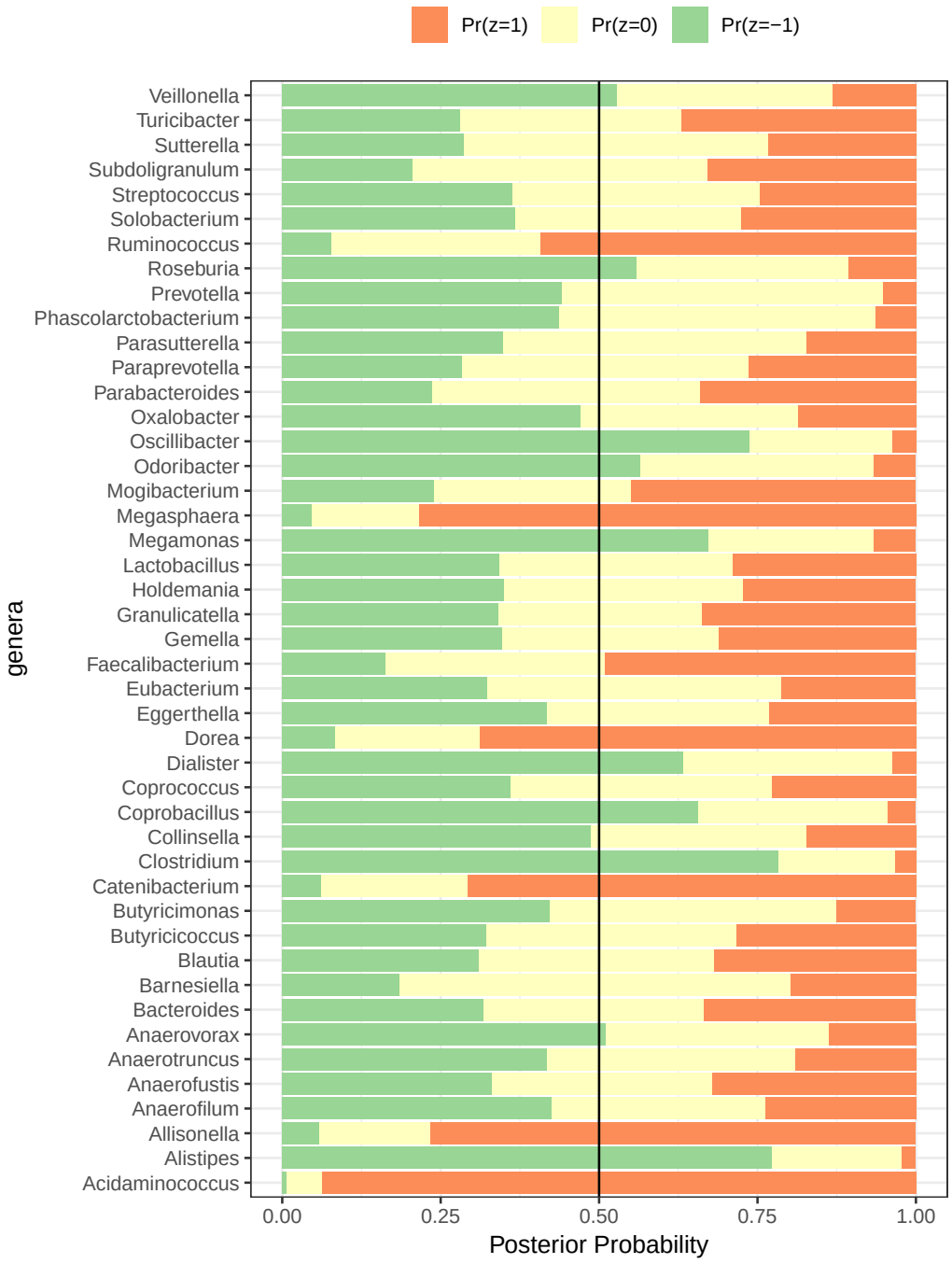
Table 2.3: Posterior inference for β for phylum level COMBO data set.

	Uniform			Sparse		
	mean	sd	95% credible interval	mean	sd	95% credible interval
β_0	24.48	2.81	(18.56,30.69)	24.62	2,75	(18.29,30.34)
β_1	0.32	0.74	(-0.99,1.99)	0.49	0.82	(-0.99,2.33)

Analyses of the COMBO data at genus level

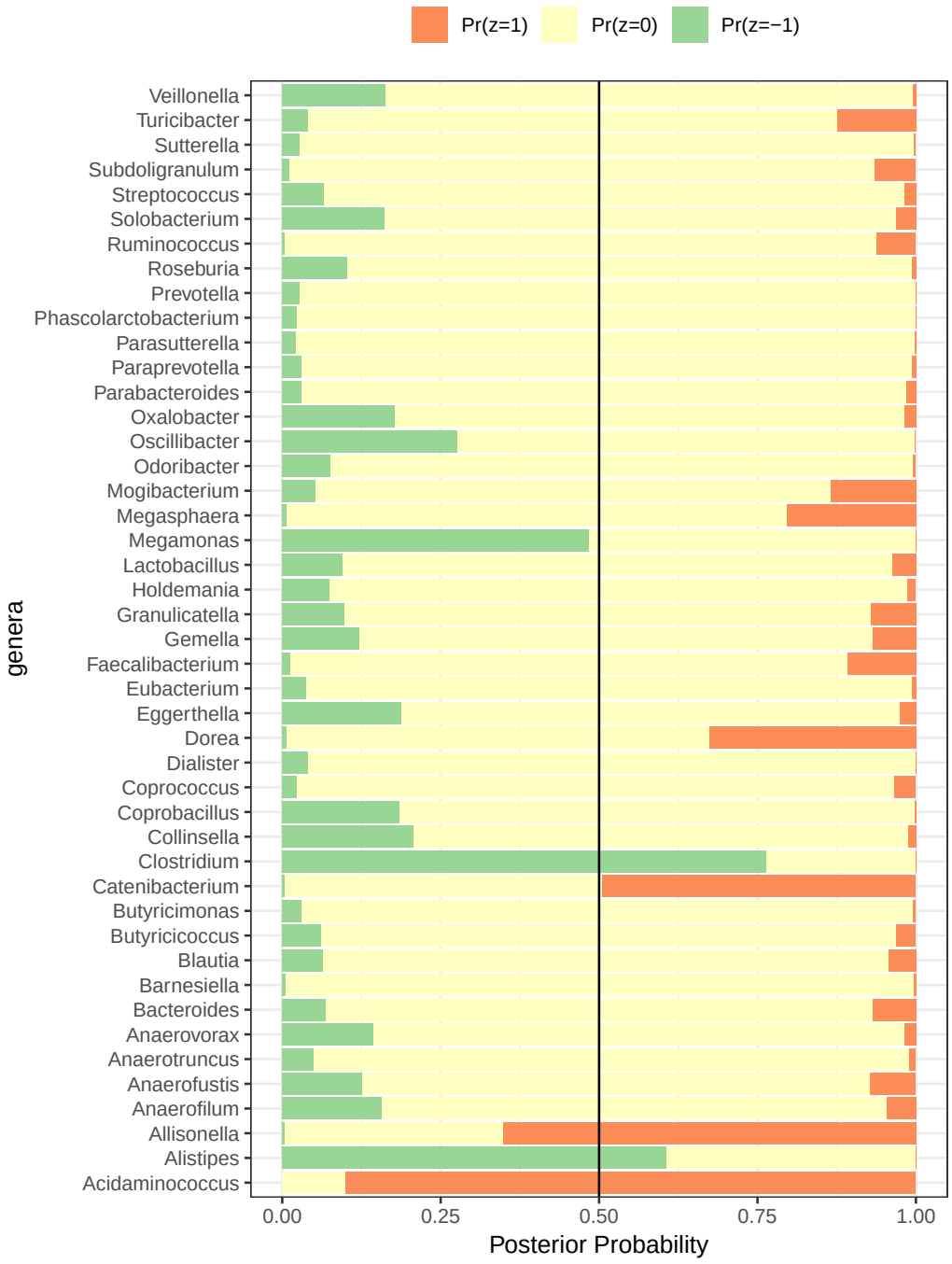
Figure 2.13, Figure 2.14, Figure 2.15 and Table 2.4 show the analysis results for the genera level COMBO data. The posterior probability plot under the sparse prior (Figure 2.14) indicate that *Acidaminococcus, Allisonella* have a probability of 0.91 and 0.63 in the z_+ set and *Alistipes, Clostridium* have a probability of 0.64 and 0.79 in the z_- set. Using 0.5 as the threshold, we conclude that under the sparse prior, balance is identified as the normalized log ratio between*Acidaminococcus, Allisonella* and *Alistipes, Clostridium*. *Acidaminococcus, Allisonella* and *Clostridium* belong to the *Firmicutes* phylum and *Alistipes* belongs to the *Bacteroidetes* phylum. This finding is consistent with the previous report that the*Firmicutes* to *Bacteroidetes* ratio is associated with diabetes and obesity.

In comparison, the uniform prior leads to more taxa being identified in the balance index and fewer taxa with a high posterior probability in the z_0 set (Figure 2.13). In particular, *Ruminococcus* (belongs to *Firmicutes*), *Megasphasera* (belongs to *Firmicutes*), *Dorea*(belongs to *Firmicutes*), *Catenibacterium* (belongs to *Firmicutes*) are identified in the numerator of balance index, whereas *Roseburia* (belongs to *Firmicutes*), *Oscillibacter* (belongs to *Firmicutes*), *Megamonas* (belongs to *Firmicutes*), *Dialister* (belongs to *Firmicutes*), *Coprobacillus* (belongs to *Firmicutes*), are identified as in the denominator, in addition to those taxa found with the sparse prior. Those additional taxa all belong to *Firmicutes*, an interesting finding, indicating genera within the same phylum may behave differently. This requires more experimental investigations by collecting data on more subjects.



(a) Uniform Prior

Figure 2.13: Posterior probabilities for each genera in the z_+, z_-, z_0 set under uniform prior for z. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.

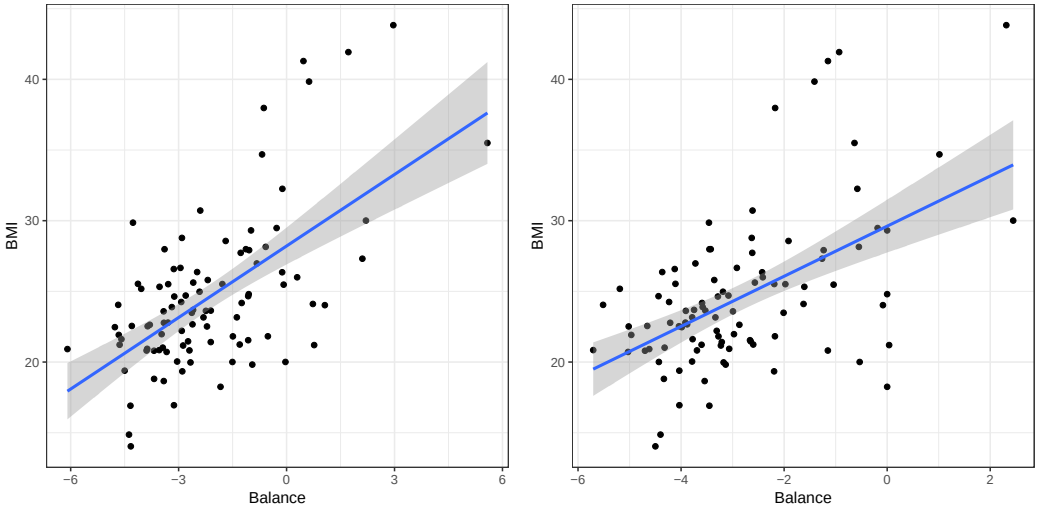


(a) Sparse Prior

Figure 2.14: Posterior probabilities for each genera in the z_+, z_-, z_0 set under sparse prior for z. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.

Table A.1 and Table A.2 present all the posterior probabilities for genera level data with uniform and sparse prior for z using 5 different starting values. The posterior probabilities are not sensitive to the initial values used in the MCMC algorithm.

Figure 2.15 shows the relationship between the outcome BMI and the estimated balance based on the 0.5 threshold cut and under two different priors for z, indicating the association between the estimated balance and BMI. Table 2.4 summarizes the posterior distributions of two coefficients. Consistent with the phylum level data, β_1 has a larger posterior mean (2.14) in the sparse prior compared to the uniform one (1.66).



(a) Uniform Prior

(b) Sparse Prior

Figure 2.15: BMI vs the estimated balance for COMBO genus level data. Blue line represents the fitted line from the ordinary least squares.

Table 2.4: Posterior inference for β with genera level data in COMBO study

	Uniform			Sparse		
	mean	sd	95% credible interval	mean	sd	95% credible interval
β_0	25.70	2.88	(19.99,31.41)	26.39	2.99	(20.14,31.89)
β_1	1.66	0.53	(0.48,2.63)	2.14	0.55	(1.12,3.28)

2.5.2. UK twin study

We also apply the proposed method on a dataset collected on twins in the United Kingdom. Goodrich et al., 2016 conducted a population based microbiome study on a large set of UK twins. In this study, extensive health data were collected on twins across the country together with gut microbiome data. The aims of the study include characterizing health status in twins on a population base and investigating the genetic/environmental effects or their interactions on general health. The study enrolled both homozygote and heterozygote twins and some subjects had multiple visits during the study. Using this data, we are interested in investigating how microbiome changes with aging. As we do not have direct measures on aging, we use actual age as a proxy outcome.

For our analyses, we first remove subjects whose microbiome data or the sibling's microbiome data are missing and remove the bacterial genera that appear in fewer than 20% of the subjects. We select one twin from each family and randomly divide the dataset into two sets of first twin and second twin. We call the first dataset 'Twin' dataset and the second one 'sibling' dataset, in order to differentiate the two. Each dataset contains 1224 subjects. A total of 65 genera that belong to 7 phyla are analyzed.

We choose the same values for the hyperparameters as in the COMBO data analyses. Similarly, we analyze the data under two priors for the balance indicator z: uniform prior and sparse prior on both the twin and the sibling datasets. We run a total of $3 * 10^5$ iterations and with $5 * 10^e$ burn-in steps.

Analyses of UK twin data at phylum level

Table 2.5 and Table 2.6 show the posterior probabilies in each of the z_+, z_-, z_0 set for the 7 phyla in the twin dataset. Five different starting values are used to run the MCMC and the posterior probabilities are similar across those starting values, indicating that the chain has converged. Figure 2.16 plots the posterior probabilities based on the fist set of the initial values.

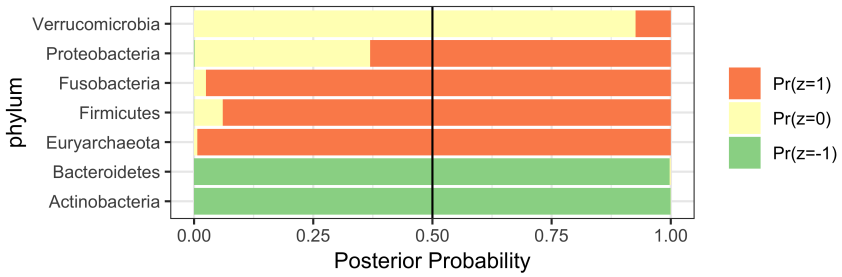
Table 2.5: Posterior probabilities with 5 starting points. Results are from phylum level twin data using uniform prior for z.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Firmicutes	0.94	0.00	0.06	0.94	0.00	0.06	0.94	0.00	0.06	0.94	0.00	0.06	0.94	0.00	0.06
Actinobacteria	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Proteobacteria	0.63	0.00	0.37	0.62	0.00	0.38	0.63	0.00	0.37	0.62	0.00	0.37	0.63	0.00	0.37
Verrucomicrobia	0.07	0.00	0.93	0.07	0.00	0.93	0.07	0.00	0.93	0.07	0.00	0.93	0.07	0.00	0.93
Bacteroidetes	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Fusobacteria	0.98	0.00	0.02	0.97	0.00	0.03	0.98	0.00	0.02	0.98	0.00	0.02	0.98	0.00	0.02
Euryarchaeota	0.99	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.01

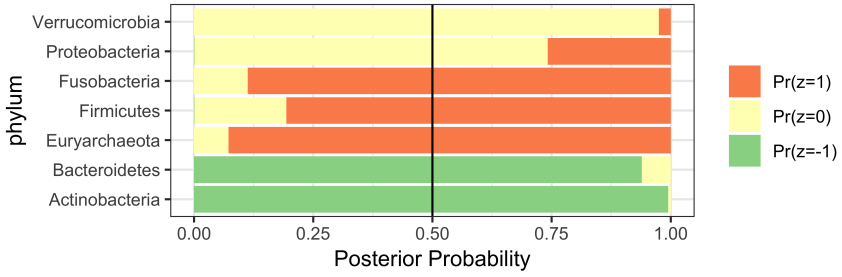
Table 2.6: Posterior probabilities with 5 starting points. Results are from phylum level twin data using uniform prior for z.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Firmicutes	0.81	0.00	0.19	0.82	0.00	0.18	0.80	0.00	0.20	0.81	0.00	0.19	0.81	0.00	0.19
Actinobacteria	0.00	0.99	0.01	0.00	1.00	0.00	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01
Proteobacteria	0.26	0.00	0.74	0.25	0.00	0.75	0.25	0.00	0.75	0.24	0.00	0.76	0.26	0.00	0.74
Verrucomicrobia	0.02	0.00	0.98	0.02	0.00	0.98	0.03	0.00	0.97	0.02	0.00	0.98	0.03	0.00	0.97
Bacteroidetes	0.00	0.94	0.06	0.00	0.94	0.06	0.00	0.94	0.06	0.00	0.94	0.06	0.00	0.94	0.06
Fusobacteria	0.89	0.00	0.11	0.89	0.00	0.11	0.88	0.00	0.12	0.88	0.00	0.12	0.89	0.00	0.11
Euryarchaeota	0.93	0.00	0.07	0.92	0.00	0.08	0.92	0.00	0.08	0.92	0.00	0.08	0.92	0.00	0.08

Comparing the uniform prior and the sparse prior for z, and using a 0.5 threshold to declare set membership, the only difference between these two priors is *Proteobacteria*. The result is consistent with our belief that under the sparse prior, balance index should contain fewer taxa. With such a prior belief, our data does not support that *Proteobacteria* is one of the components in balance index. We conclude that the balance is composed of the scaled log ratio among *Fusoaacteria*, *Firmicutes*, *Euryarchaeota* and *Bacteroidetes*, *Actinobacteria*.



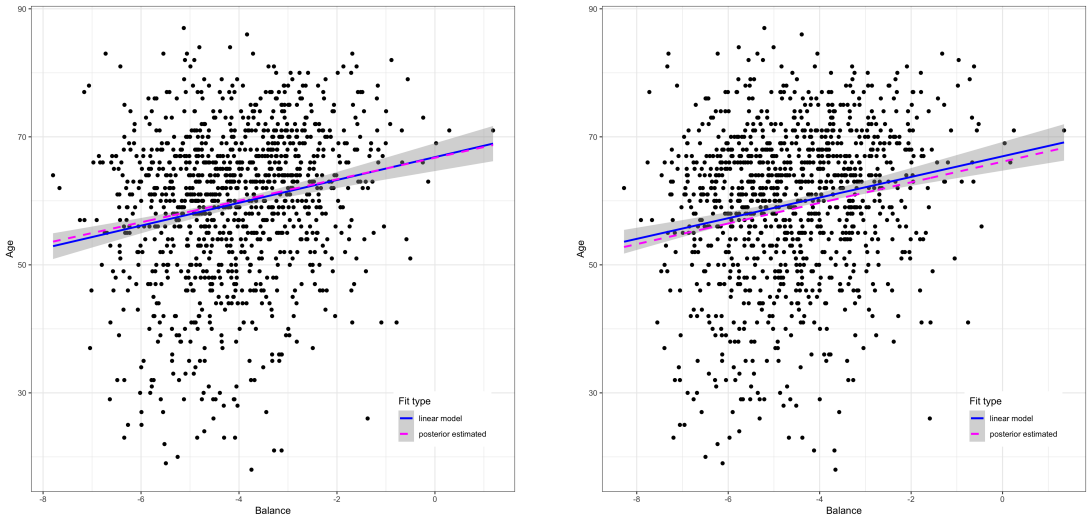
(a) Uniform Prior



(b) Sparse Prior

Figure 2.16: Posterior probabilities for individual phylum being in each of the z_+, z_-, z_0 sets for the UK twin data. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.

Figure 2.17 summarizes the relationship between age and estimated balance using 0.5 as a threshold. A summary of the posterior distribution of $\boldsymbol{\beta}$ is shown in Table 2.7, indicating the the balance identified is indeed associated with age. The posterior mean of $\boldsymbol{\beta}$ is similar under two different priors, but the variance is larger with the sparse prior indicating the posterior distribution of $\boldsymbol{z}$ is more flat compared to the uniform prior.



(a) Uniform Prior

(b) Sparse Prior

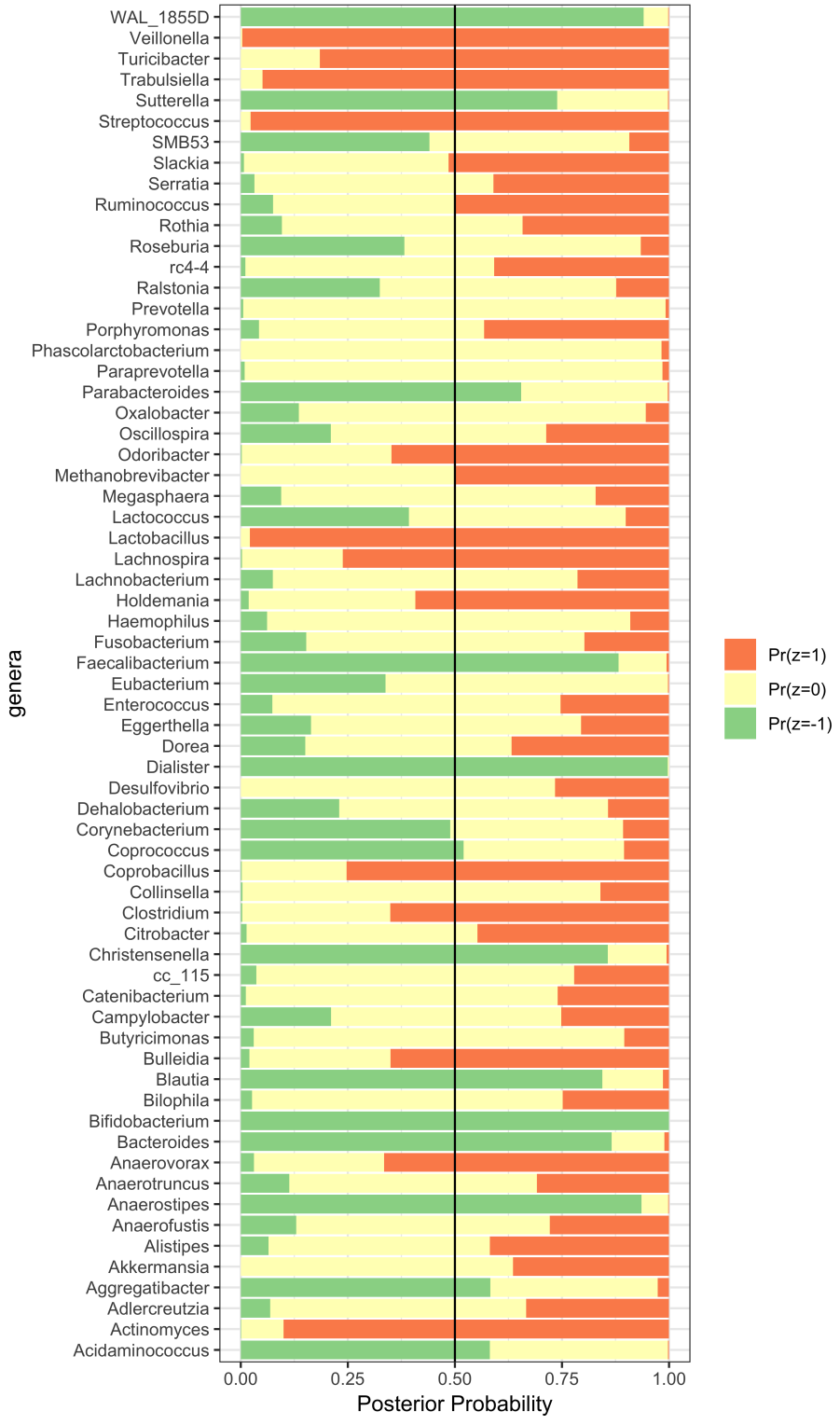
Figure 2.17: Age vs the estimated balance in UK twin data. Blue line represents the fitted line from the linear least squares. Magenta line represents the posterior mean of β

Table 2.7: Posterior inference for β with phylum level data in UK twin study

	Uniform			Sparse		
	mean	sd	95% credible interval	mean	sd	95% credible interval
β_0	66.73	1.39	(63.46,69.22)	66.12	2.82	(56.25,69.62)
β_1	1.68	0.24	(1.11,2.23)	1.61	0.39	(0.94,2.62)

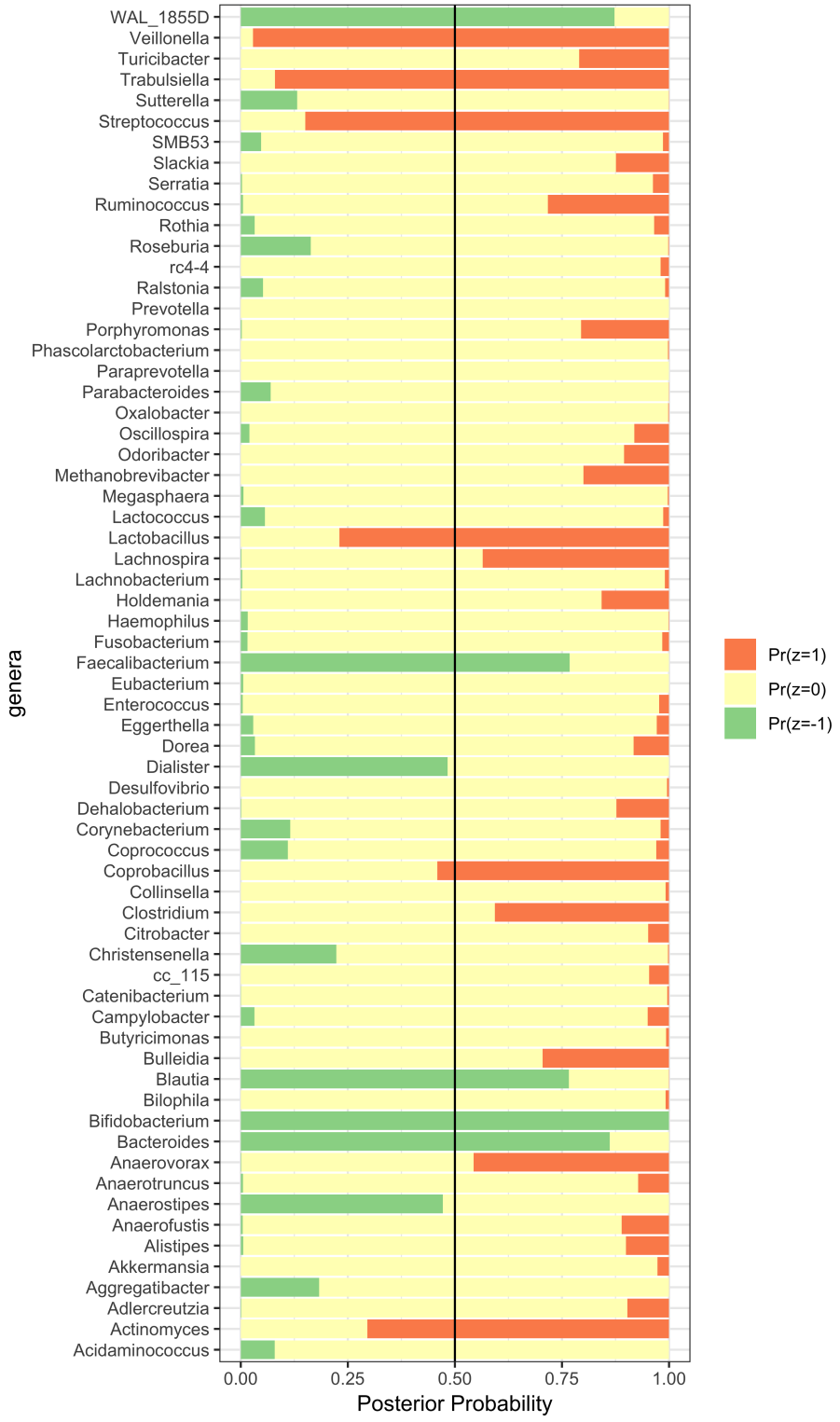
Analyses of UK twin data at genus level

Tables A.3 and Table A.4 show the posterior probabilities of a genus being in each of the z_+, z_-, z_0 sets for the 65 genera in the twin dataset. Five different starting points are used to run the MCMC and the posterior probabilities are similar across those starting points, indicating that the chain has converged. Figure 2.18 and Figure 2.19 plot the posterior probabilities based on the first set of initial values. The posterior probabilities for genera are different under uniform and sparse priors. However, certain genera are consistently identified as being the components of the balance. Given very large sample space of the latent indicator of 65 dimensions, the sparse prior should be used.



(a) Uniform Prior

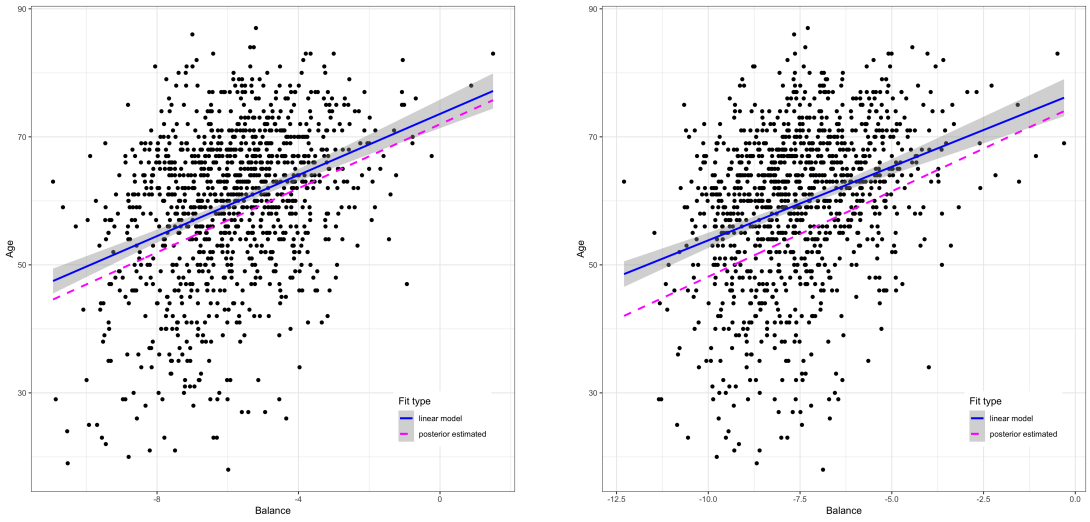
Figure 2.18: Posterior probabilities for individual genus being in each of the z_+, z_-, z_0 sets for the UK twin data under uniform prior for $\boldsymbol{z}$. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.



(a) Sparse Prior

Figure 2.19: Posterior probabilities for individual genus being in each of the z_+, z_-, z_0 sets for for the UK twin data under sparse prior for z. Green bar represents the posterior probability in z_- set; Red bar represents the z_+ set. The vertical line is the 0.5 cutoff line.

Figure 2.20 shows the plots of the estimated balance and age under two different priors, indicating similar trend of association. Table 2.8 shows the posterior mean of β_1, which are close under two different prior assumptions.



(a) Uniform Prior

(b) Sparse Prior

Figure 2.20: Age vs the estimated balance for UK twin data at the genus level. Blue line represents the fitted line from the linear least squares. Magenta line represents the posterior mean of $\boldsymbol{\beta}$

Table 2.8: Posterior inference for $\boldsymbol{\beta}$ under uniform and sparse prior for $\boldsymbol{z}$ for UK twin data at the genus level.

	Uniform			Sparse		
	mean	sd	95% credible interval	mean	sd	95% credible interval
β_0	71.97	4.04	(64.14,79.96)	74.83	5.05	(65.2,84.86)
β_1	2.50	0.27	(1.98,3.06)	2.67	0.34	(2.05,3.39)

2.6. Discussion

In this Chapter, we have proposed a Bayesian balance-regression and a MCMC stochastic search algorithm to identify the microbial signature, termed balance, which is related to the outcome of interest. The balance signature reflects the mutualistic or repellent interactions among microbes. Balance can be regarded as the extension of the log ratio between two groups of microbes that are highly indicative of healthy/pathological traits. The proposed method automatically selects the

most relevant reference taxa in the denominator of the balance, unlike other approaches that adopt additive log ratio or central log ratio transformation, and at the same time also preserves the scale-invariance as well as sub-compositional coherence properties. Results from the analysis of the association between gut microbiome and BMI and IBD demonstrated the model and methods can indeed identify the balance feature that is closely associated with these outcomes. The balance can be used as an index for dysbiosis of the microbial community,

The MCMC stochastic search in our method allows feasible exploration of the moderate to high dimensional space (3^p-dimensional for p taxa) whereas the search algorithm Rivera-Pinto et al., 2018 only searches for a small space of possible balance since it only allows adding one component to the positive part or the negative part of the previous sets during the iterations. Our algorithm allows the elements in the previous sets being removed out of the balance set or switching from the positive set to the negative set or vice versus. Another advantage of using a Bayesian approach to balance-regression is that it automatically provides an assessment of uncertainty of the balance identified and its effect on the outcome.

As for any Bayesian methods, one needs to choose sensible values for the hyperparameters in the prior distributions. In our analysis, we used non-informative or vague priors. However, in studies where sample size is small or noise level is high, the selection result can be affected by the prior parameters. During MCMC sampling, we obtain the the posterior inclusion probabilities of all three sets for each of the taxa. We use the posterior modes to choose the final set of z and the corresponding balance. Since the conditional posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{z}}$ has a closed form, statistical inference about $\boldsymbol{\beta}_{\boldsymbol{z}}$ can be made based on the final estimate of $\mathbf{z}$. In this case, our inference of $\boldsymbol{\beta}_{\boldsymbol{z}}$ can be interpreted as conditioning on z. However, in real data analysis, the posterior probabilities in the three sets might only differ slightly in magnitudes, which makes posterior inference of z difficult. In this case, the posterior inference about $\boldsymbol{\beta}_{\boldsymbol{z}}$ can be performed on the sampling points obtained during each iteration of MCMC in the framework of model average.

Our proposed balance-regression model and the MCMC method are very different from the Bayesian variable selection in linear models (Holmes and Held, 2006). One key difference is that the covariate (i..e, the balance) in our model (2.2) is unknown and has to be identified. Our models have only one covariate, but it is not observed. In constrast, the standard Bayesian linear model selection aims to identify the relevant covariates from a set of fixed covariates. The proposed method aims to

discover a particular form from a subset of taxa that act as one explanatory entity for the outcome. In this regard, we treat the regression coefficient as a nuisance parameter and integrate it out of the target distribution. Whereas in Bayesian variable selection, whether a covariate is in the final model depends on the associated regression coefficient being zero or not.

Identifying the balance that is associated with an clinical outcome is a challenging problem. Although we have demonstrated through simulations that the MCMC algorithm is effective in identifying the structure of the balance, our results are based on simulating the data from the proposed models. It would be interesting to further investigate the methods by simulating data from alternative models such as the compositional linear or generalized linear models studied in Lu, Shi, and Li, 2019; Shi, Zhang, and Li, 2016 for microbiome studies. In fact, if the balance indicator vector z is known, the balance-regression model proposed in this paper is a special case of the model in Shi, Zhang, and Li, 2016, where the positive coefficients of all relevant taxa are assumed to be the same and the negative coefficients of all relevant taxa are also assumed to be the same.

CHAPTER 3

BAYESIAN PROBIT BALANCE REGRESSION

3.1. Introduction

Interrogating gut microbiome provides an important tool to understand not only the gastrointestinal tract related diseases, but also abnormalities in immune system, respiratory system and even in the brain. How to determine the association between each taxon in the microbiome and the disease status has been a challenging problem due to both high dimensionality and lack of measurements of absolute bacterial abundances. The $p >> n$ problem in microbiome data analysis can be mitigated using high-dimensional statistical methods (Li, 2015; Lu, Shi, and Li, 2019; Shi, Zhang, and Li, 2016), and satisfactory results have been achieved in numerical simulations as well as in real data analysis. To reduce the complexity in interpretation of results and to provide an intuitive yet accurate inference, a simplified summary of microbiome compositional data is necessary.

Finding simple summary of microbiome compositional data that is biologically relevant is to some extent equivalent to dimension reduction in statistics, In the most extreme case, we can combine all the dimensions and summarize them into one number. Due to the skewness of microbiome compositional data, a geometric mean is more appropriate than a simple average. However, the geometric mean tells us nothing about the relationship among different taxa. In healthy human, although its composition might be different, the microbiota community roughly maintains a homeostasis and is resistant to slight or moderate perturbations (Das and Nair, 2019). Dysbiosis, though not fully characterized, is often associated with various disease. Thus it is important to simplify the microbiome data with a number that can capture the interplay within the microbiota.

As microbiome data is compositional, three principles, namely permutation invariance, scale invariance and subcompositional cohesion must be satisfied (Aitchson, 2003). With this regard, a log ratio transformation is a natural choice to transform the compositional data (Aitchson, 2003). The concept of balance introduced by Rivera-Pinto et al., 2018 is a useful index to characterize the microbiome data with a single scalar, which meets the three principles and also captures the relationship within the microbiota. The balance defined as in equation (2.1) in Chapter 2, can be regarded as a weighted log ratio between two partitions of a compositional vector, say $\mathbf{x} = (x_1, x_2, \cdots, x_p)$.

Let $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ be a p-dimensional indicator vector with three values, 1, -1, and 0, where $z_i = 1$ indicates that the corresponding part in the composition x_i is in the numerator of a ratio, $z_i = -1$ indicates the denominator; and $z_i = 0$ indicates the parts that are not a member of the balance. Let z_+, z_-, and z_0 be non-overlapping sets of indices of the elements in $\mathbf{z}$ whose values are 1, -1 and 0, respectively, and $z_+ \cup z_- \cup z_0 = \{1, 2, \ldots, p\}$. Given $\boldsymbol{z}$, the corresponding balance of $\mathbf{x}$ is defined as

$$B_z = \sqrt{\frac{m_+ m_-}{m_+ + m_-}} \log \frac{\left(\prod_{i \in z_+} x_i\right)^{1/m_+}}{\left(\prod_{i \in z_-} x_i\right)^{1/m_-}}, \tag{3.1}$$

where m_+, m_-, m_0 denotes the number of indices in z_+, z_-, z_0 respectively.

In this chapter, we extend Bayesian linear Balance regression in Chapter 2 to the binary outcomes such as disease status. In Bayesian literature, a popular method for modeling a binary outcome is to use auxiliary continuous outcome variable (Albert and Chib, 1993; Holmes and Held, 2006) that can be modeled with a linear model. The final binary outcome is determined by the sign of this auxiliary variable. This setup leads to closed-form expressions of posterior distribution and conditional distributions. Mathematically, it corresponds to a generalized linear model with probit link and computationally such a setup makes the sampling procedure faster (Holmes and Held, 2006)

3.2. Bayesian balance probit regression

For a binary outcome, let y_i be the outcome for subject i that takes only two possible values 0 and 1 and $\mathbf{x}_i$ be the p-dimensional compositional vector of p bacterial relative abundances. Similar to model (2.2), we define $\mathbf{B_{1z}} = (\mathbf{1}, \mathbf{B}_z)$ to be the $n \times 2$ matrix with columns of $\mathbf{1}$ and the vector $\mathbf{B}_z$, where $\mathbf{z}$ is the indicator vector that defines the balance and B_{zi} is defined as in equation (3.1). The following model with a hidden variable y_i^* corresponds to the generalized linear model with a probit link (Holmes and Held, 2006),

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ y_i^* | \boldsymbol{z} &= \beta_{0z} + \beta_{1z} B_{zi} + \epsilon_{iz} \\ \epsilon_{iz} &\sim N(0, 1) \end{aligned} \tag{3.2}$$

The prior distribution for $\boldsymbol{\beta}_z = (\beta_{0z}, \beta_{1z})$ is chosen as the multivariate normal with mean $\boldsymbol{b_0}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$, whose entries are v_0, v_1. After integrating out $\boldsymbol{\beta}$, the conditional distribution of $(\mathbf{y}, \mathbf{y}^*)|\mathbf{z}$ is

$$\frac{|\boldsymbol{\Sigma^*}|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}^* - \mathbf{B_{1z}}\mathbf{b_0})^T \boldsymbol{\Sigma^*}(\mathbf{y}^* - \mathbf{B_{1z}}\mathbf{b_0})\right\} \times g(\mathbf{y}|\mathbf{y}^*) \tag{3.3}$$

with $\boldsymbol{\Sigma^*} = (\mathbf{I} + \mathbf{B_{1z}}\boldsymbol{\Sigma}\mathbf{B_{1z}}^T)^{-1}$, and $g(\mathbf{y}|\mathbf{y}^*)$ is an indicator function which truncates the multivariate normal distribution of $\mathbf{y}^*$ to the appropriate region. Therefore $\mathbf{y}^*|\mathbf{y}, \mathbf{z}$ simply follows a truncated normal distribution with truncation at zero. Sampling from this distribution is achieved by Gibbs sampling the multivariate truncated normal distribution, $y_i^*|y_{-i}^*, y_i, \mathbf{z}$, where y_{-i}^* denotes the auxiliary variable $\mathbf{y}^*$ with the ith variable removed. This is a univariate truncated normal distribution with the mean and variance that can be obtained from the leave-one-out marginal predictive density derived from (3.3).

The posterior distribution of $\boldsymbol{z}$ is proportional to multiplying equation (3.3) by the prior of $\mathbf{z}$ given in equation (2.3). Due to unobserved $\mathbf{y}^*$, we need to sample both $\mathbf{z}$ and $\mathbf{y}^*$ in our algorithm. The information from the observed data $\mathbf{y}$ is used to infer $\mathbf{y}^*$, which in turn determines $\mathbf{z}$. As a result we propose the following MCMC sampling steps (see Algorithm 2).

Algorithm 2: MCMC algorithm for the Bayesian probit balance-regression analysis

1 sample $\mathbf{y}^*|\mathbf{y}, \mathbf{z}$ from truncated normal whose original form is equation (3.3) and truncation occurs at 0. This is achieved by Gibbs sampling based on the density of (3.3).
2 sample $\mathbf{z}|\mathbf{y}^*$. This step is equivalent to the sampling in Chapter 2
3 return to the first step until convergence.

The posterior inference of the balance structure is the same as the Bayesian linear balance regression studied in Chapter 2. Here, we propose to carry out the predictive posterior inference for $\boldsymbol{\beta}_z$ conditional on $\mathbf{y}^*, \mathbf{z}$, as the information contained in $\mathbf{y}$ is passed onto $\mathbf{y}^*$. In each MCMC iteration, given $\mathbf{z}$ and $\mathbf{y}^*$, the posterior distribution of $\boldsymbol{\beta}_z$ is a multivariate normal with mean vector $\tilde{\boldsymbol{\mu}}$ and variance covariance matrix $\tilde{\boldsymbol{\Sigma}}$ given as

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}(\mathbf{B}_{1z}^T\mathbf{y}^* + \boldsymbol{\Sigma}^{-1}b_0),$$
$$\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}^{-1} + \mathbf{B}_{1z}^T\mathbf{B}_{1z})^{-1}.$$

From this, we can obtain the mean of the predictive posterior distribution of $\boldsymbol{\beta}_z$ given $\boldsymbol{z}$.

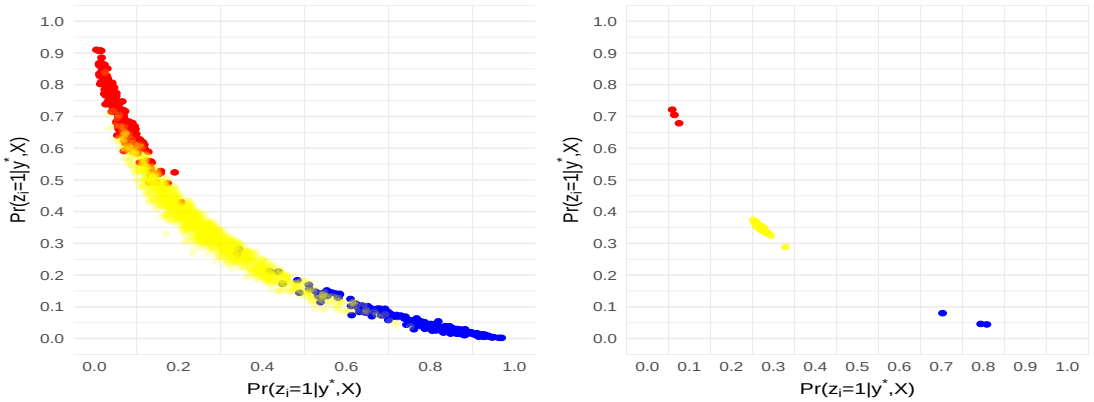
3.3. Numerical studies

The performance of Bayesian balance probit regression is evaluated with simulations. The hyperparameters in prior distributions are chosen to be uninformative or a flat prior, allowing data to take more weights in the posterior distributions. The values of the hyperparameters are summarized as the following: $b_{00} = 0, b_{01} = 0, v_0 = 10^3, v_1 = 10^3, w_1 = w_2 = 1/3$. These choices of the hyperparameters represent noninformative priors, giving the data a better chance to influence the posterior. For binary outcome, we set $c = 10^3, \quad h = 10^3$ due to the restrictions in Gibbs sampling of the latent vector $\mathbf{y}^*$. The probabilities for each taxon being in the z_+, z_-, z_0 are assumed to be equal. We set $v = 3, \lambda = 1$, corresponding to a prior estimate for σ^2 being 1. In practice, λ can be chosen as the sample variance of the outcomes, leading to an empirical prior.

For all the simulations, we use $n = 100, \quad p = 30, \quad \beta_0 = 1, \quad \sigma^2 = 1$. The balance is composed of the first three taxa in the z_+ set and the next three in the z_- set, all the rest are in the z_0 set. We evaluate strong, moderate, small and null effect size of β_{1z}, namely 1, 0.5, 0.1 and 0. To generate the compositions, we first generate n independent vector of dimension p from a multivariate normal distribution with zero mean vector and covariance of first order auto-correlation. The correlation parameter is set to 0.2 and the variance equals 10. The taxa count matrix is created by exponentiating and then applying the floor function to get the greatest smaller integer. Zero counts are replaced with 0.5 and the count matrix is normalized per row and log-transformed to get $\mathbf{X}$. The $\mathbf{y}^*, \mathbf{y}$ are generated according to balance probit regression model (3.2). Starting values for $\mathbf{z}$ is random. The mixing property of Markov chains are evaluated with several starting values of $\mathbf{z}$.

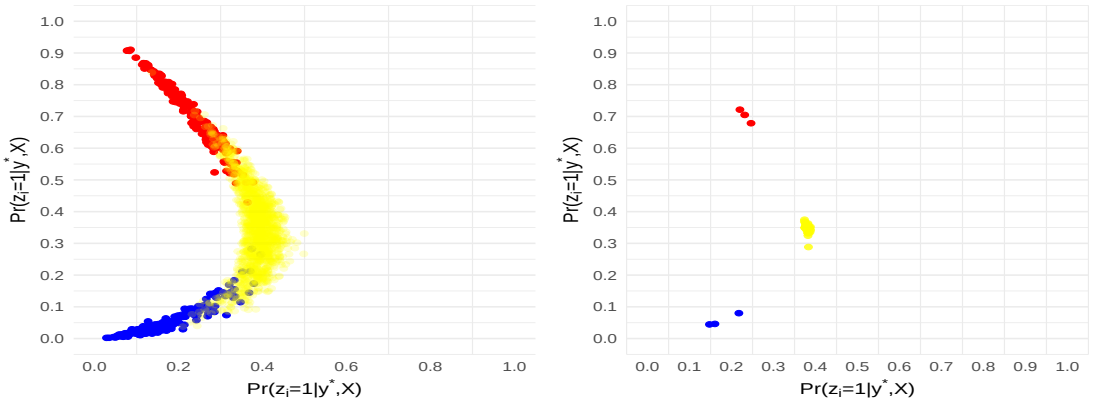
3.3.1. Results under uniform prior of $\boldsymbol{z}$

Compared to Bayesian linear balance regression, the posterior probability in the probit regression (Figures 3.1,3.2,3.3, 3.4) tends to shrink to the prior with greater degrees. The extra randomness caused by sampling the latent vector $\mathbf{y}^*$ has greater effects on the posterior inference for $\mathbf{z}$. Similar to the linear regression and as expected, stronger balance effect has a better chance of identifying the correct structure of the balance. In the worst case of weak or null effect, where $\beta = 0.1$ or $\beta = 0.0$, it is not possible to classify a taxon into any set of z_+, z_-, z_0 with a high probability.



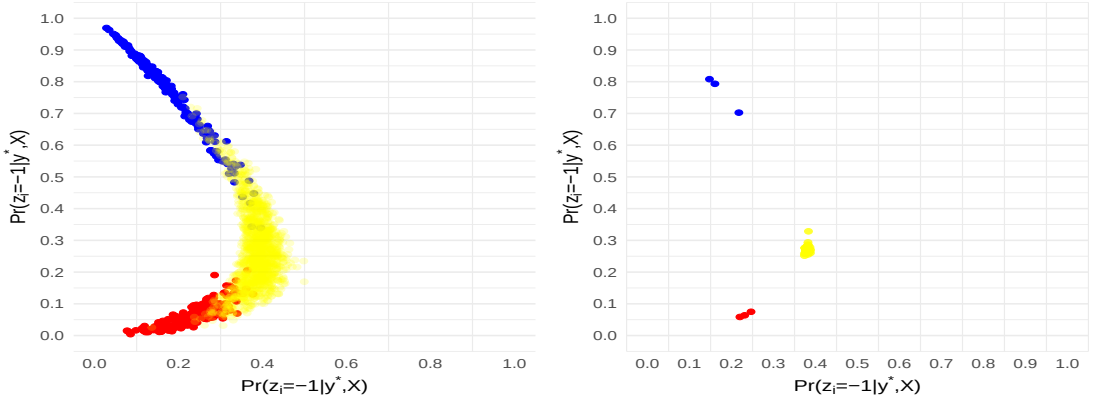
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

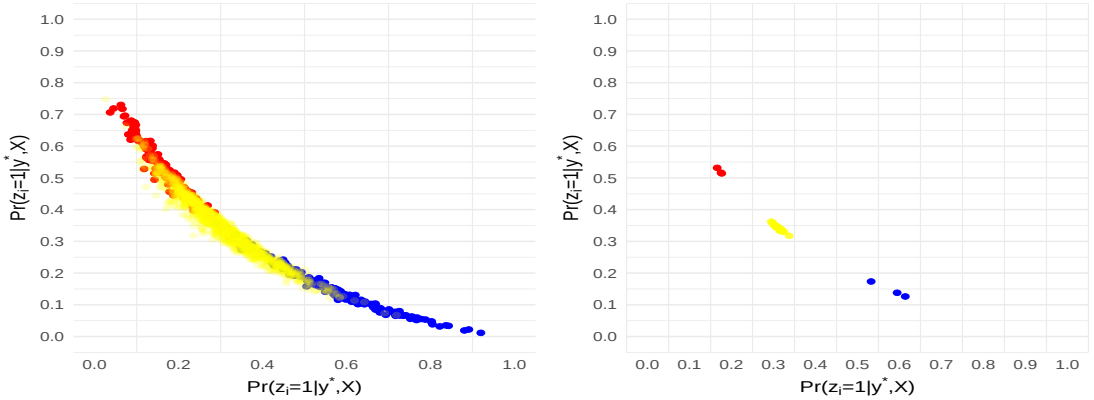
(d) averaged across simulations



(e) all simulation points

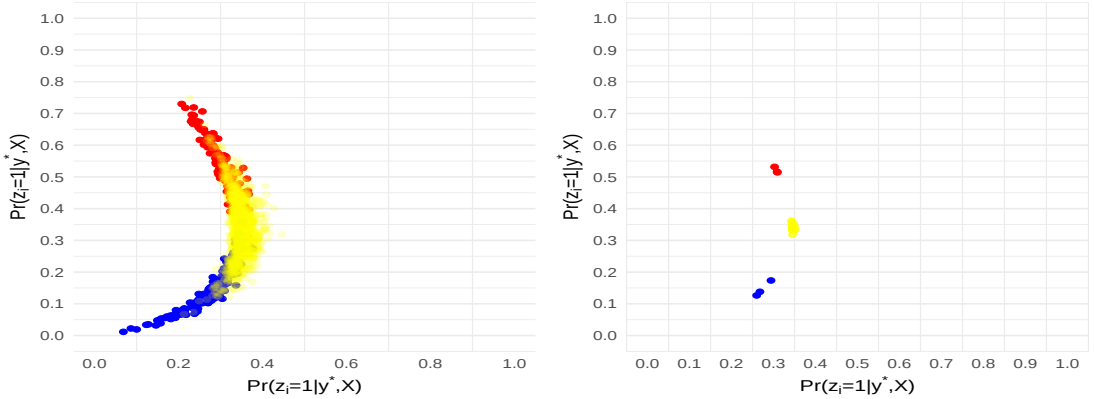
(f) averaged across simulations

Figure 3.1: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 1$ under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



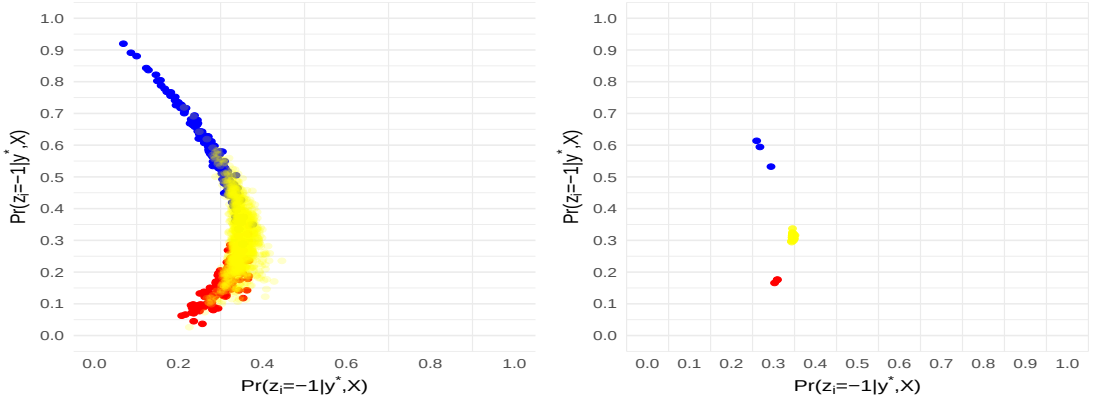
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

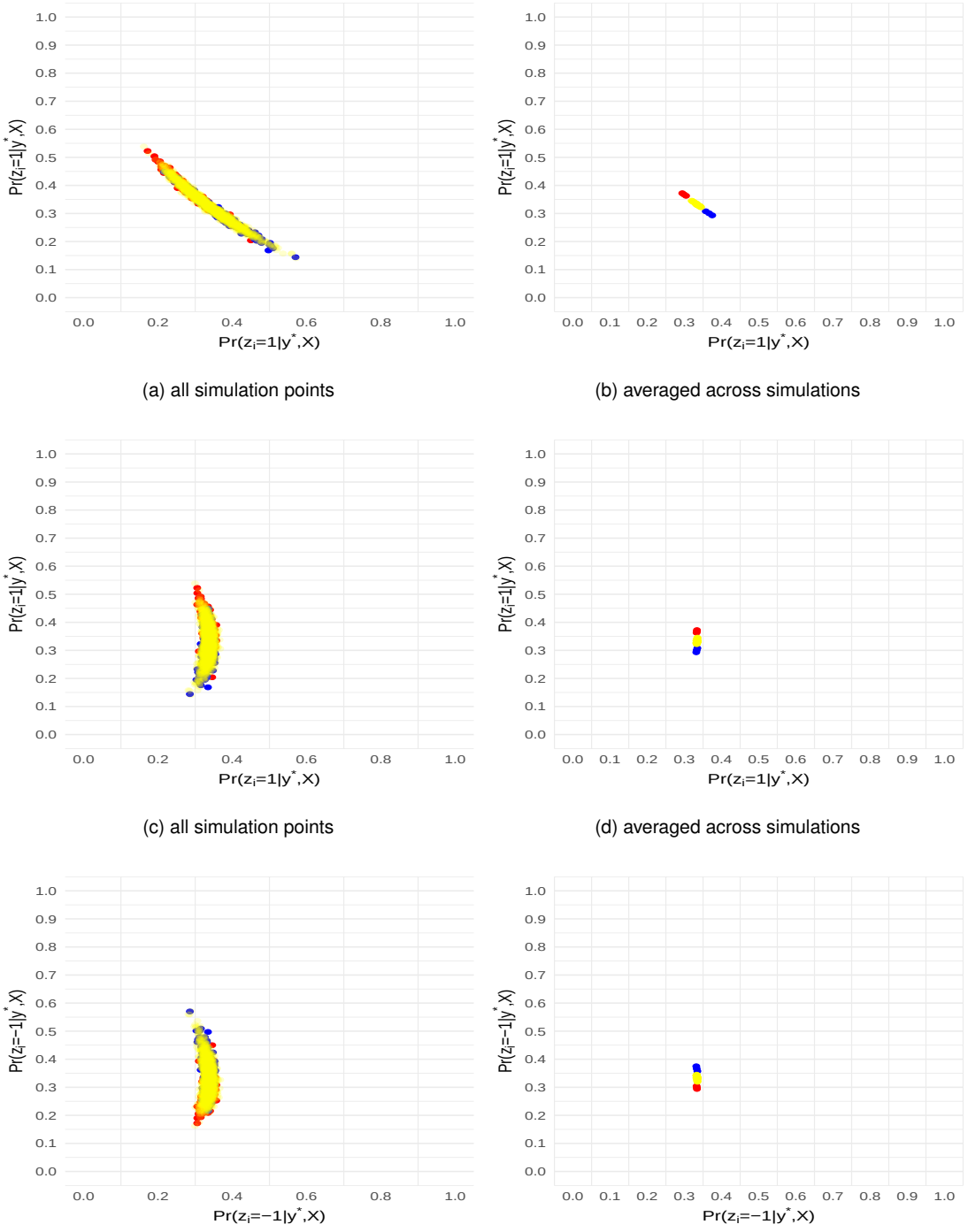
(d) averaged across simulations



(e) all simulation points

(f) averaged across simulations

Figure 3.2: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0.5$ under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



(a) all simulation points

(b) averaged across simulations

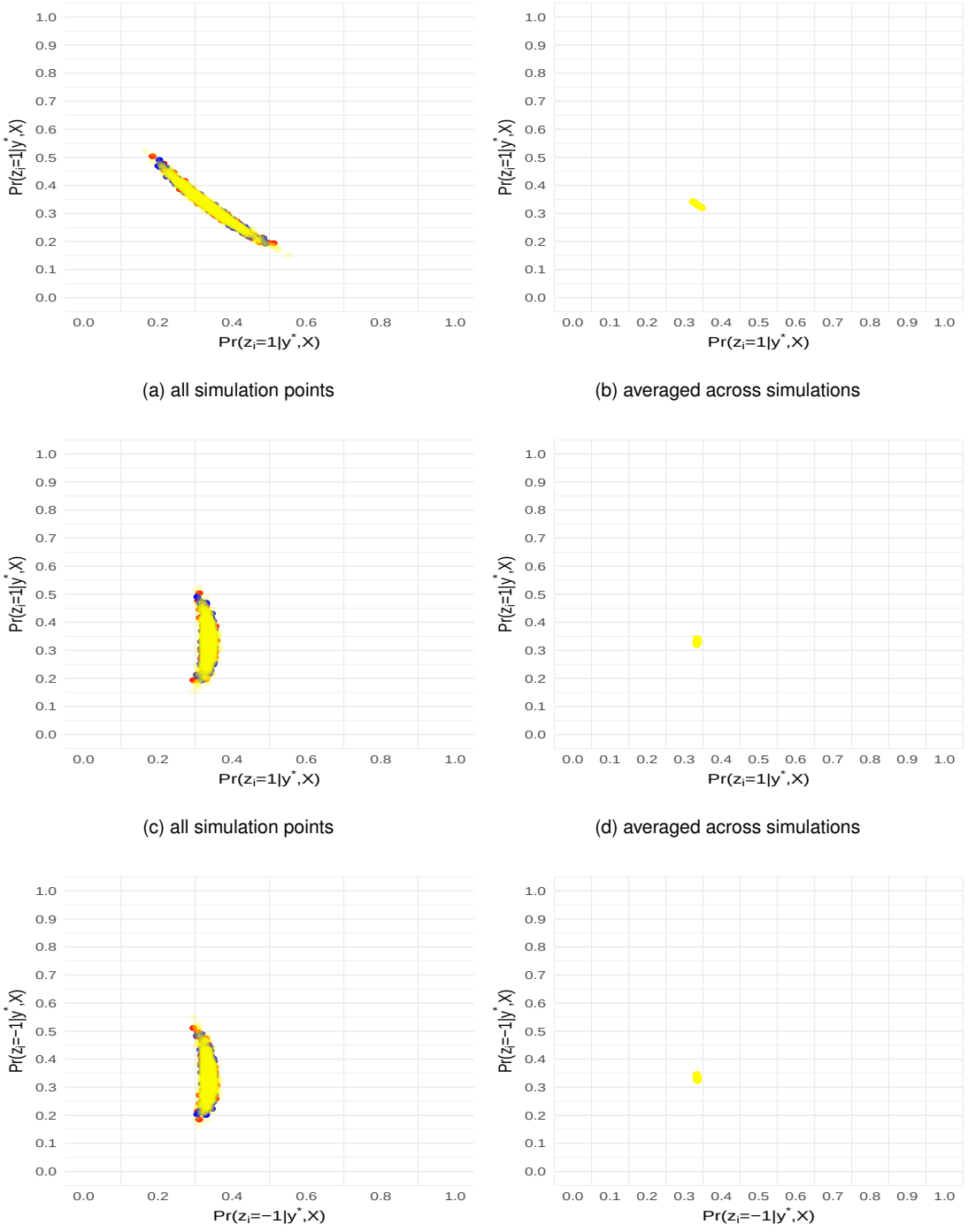
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

(f) averaged across simulations

Figure 3.3: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0.1$ under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



(a) all simulation points

(b) averaged across simulations

(c) all simulation points

(d) averaged across simulations

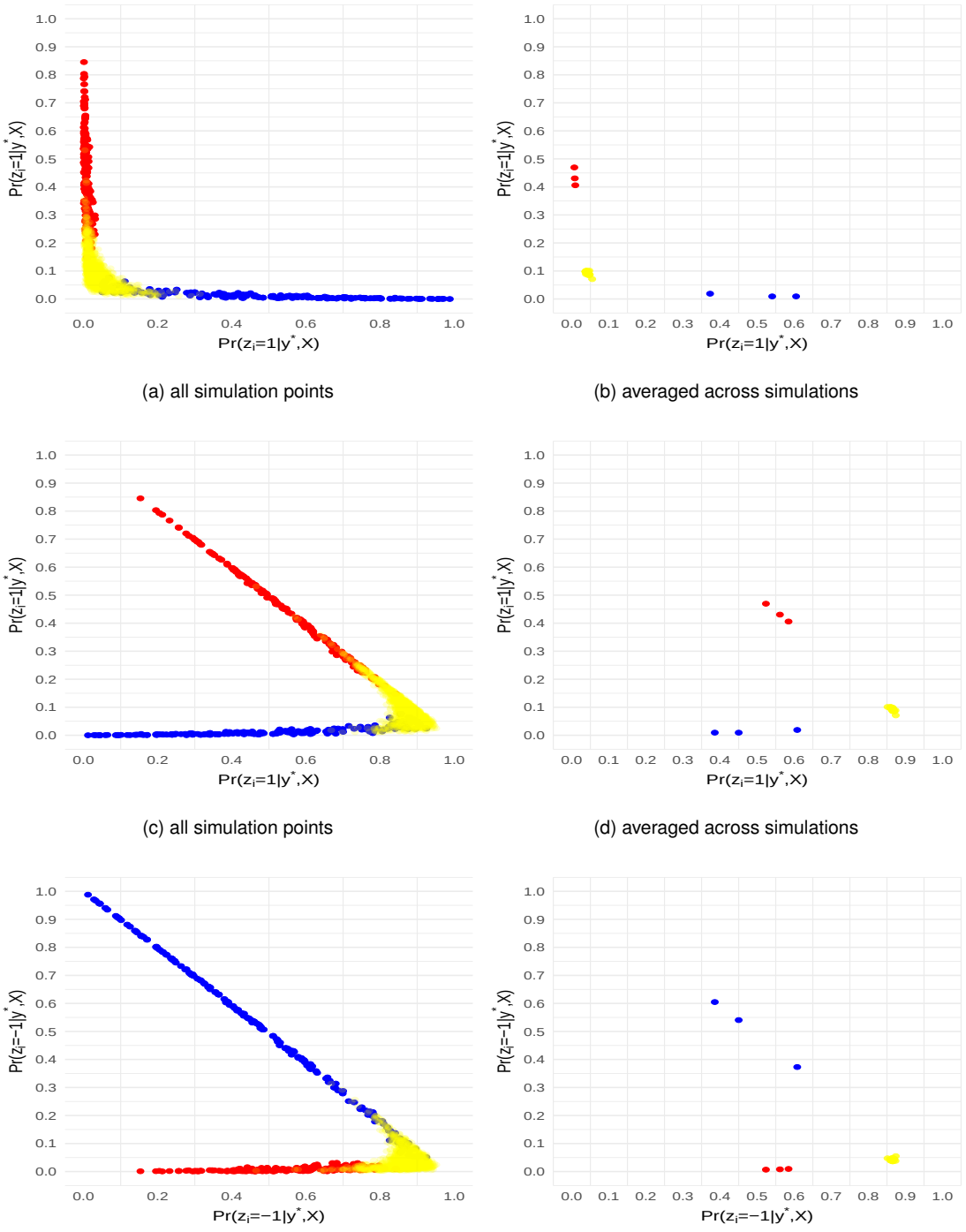
(e) all simulation points

(f) averaged across simulations

Figure 3.4: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0$ under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).

3.3.2. Results under sparse prior of z

Figures 3.5, 3.6,3.7 and Figure 3.8 summarize the posterior probabilities under sparse prior all 30 covariates across 100 simulations. Similar to the uniform prior, when the effect size is strong $\beta_{1z} = 1$ and medium $\beta_{1z} = 0.5$, all taxa can be correctly identified in the balance structure. When the effect size is small $\beta_{1z} = 0.1$ and none $\beta_{1z} = 0$, averaged posterior probabilities for each taxon shrink toward the prior and almost all of them are identified in the z_0 set. Compared to uniform prior, all the simulation points are more dispersed. With a sparse prior the shrinkage effect toward the prior distribution is even more severe in the first six components where the true value is in the z_+ or z_- sets. With a small effect size $\beta = 0.1$, almost all components have a high posterior probability of being in the z_0 set. These results show that for the Bayesian probit balance regression, taking uniform prior of the balance indicator can result in better identification of the outcome associated balance.



(a) all simulation points

(b) averaged across simulations

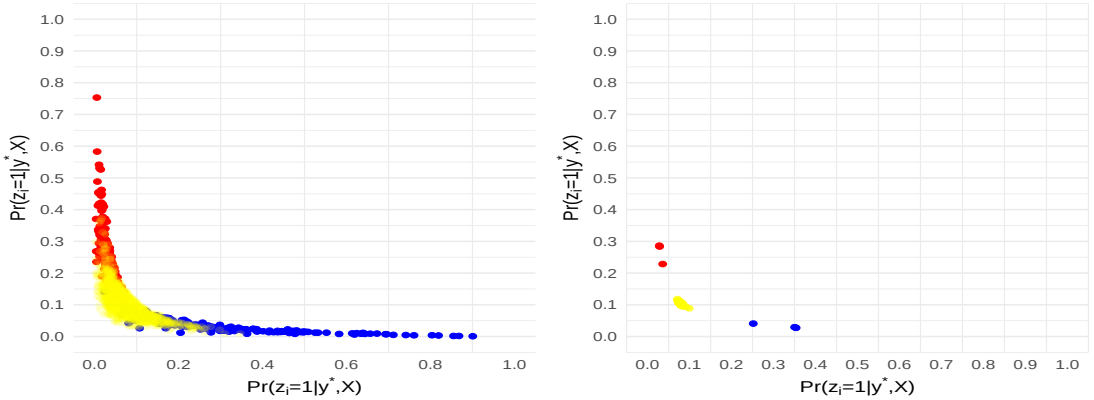
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

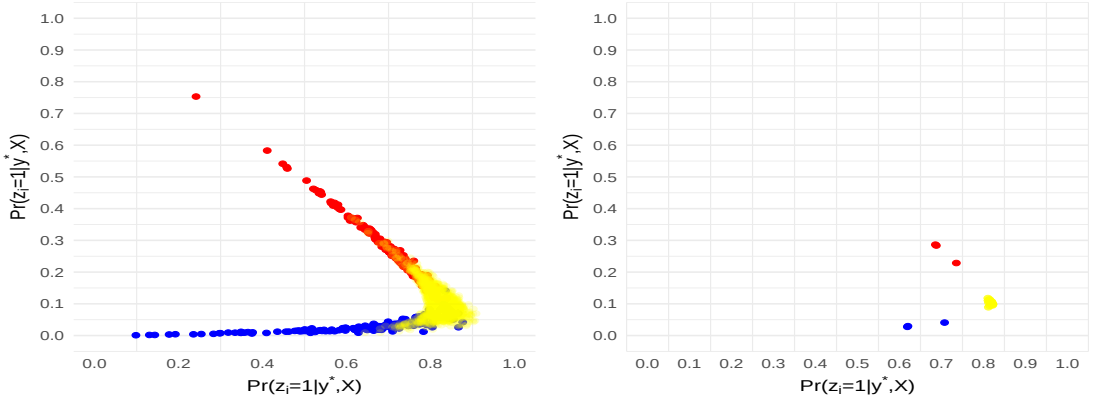
(f) averaged across simulations

Figure 3.5: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 1$ under sparse prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



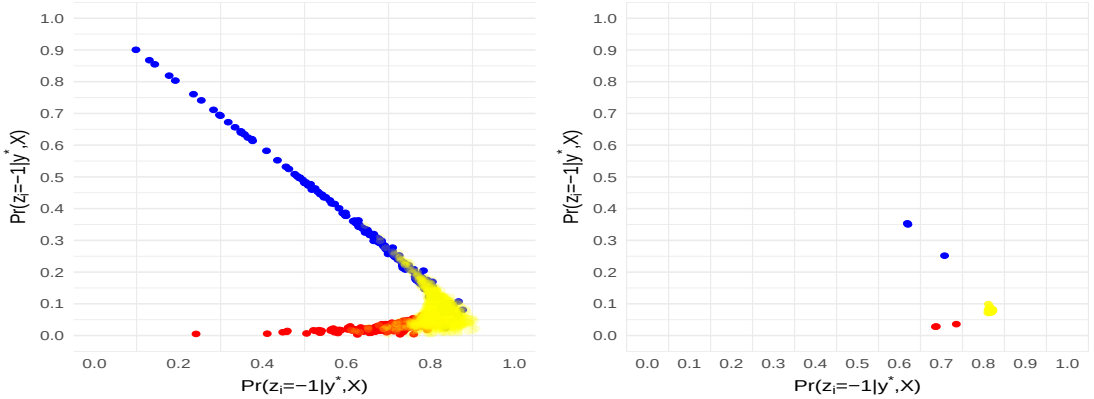
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

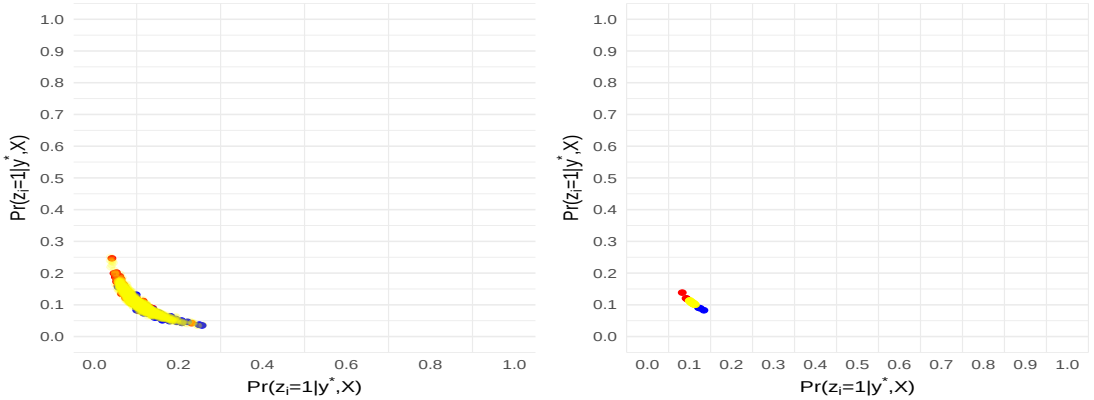
(d) averaged across simulations



(e) all simulation points

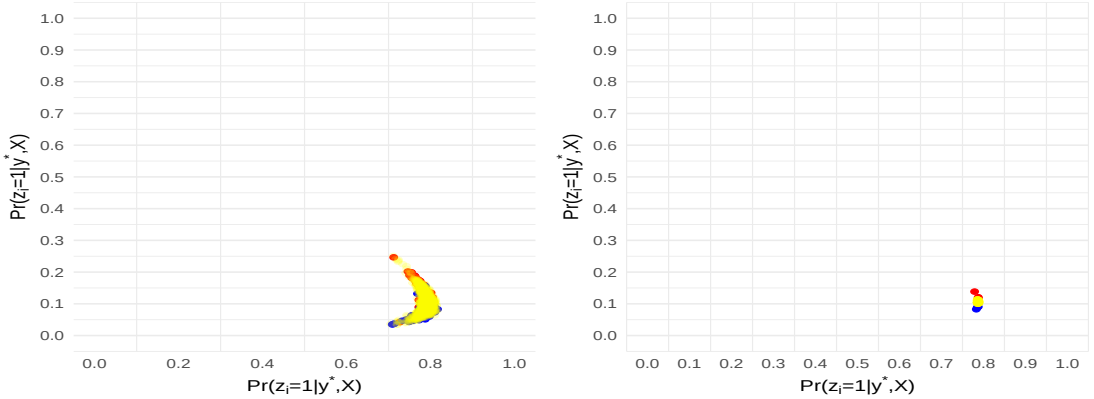
(f) averaged across simulations

Figure 3.6: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0.5$ under sparse prior for $\boldsymbol{z}$. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



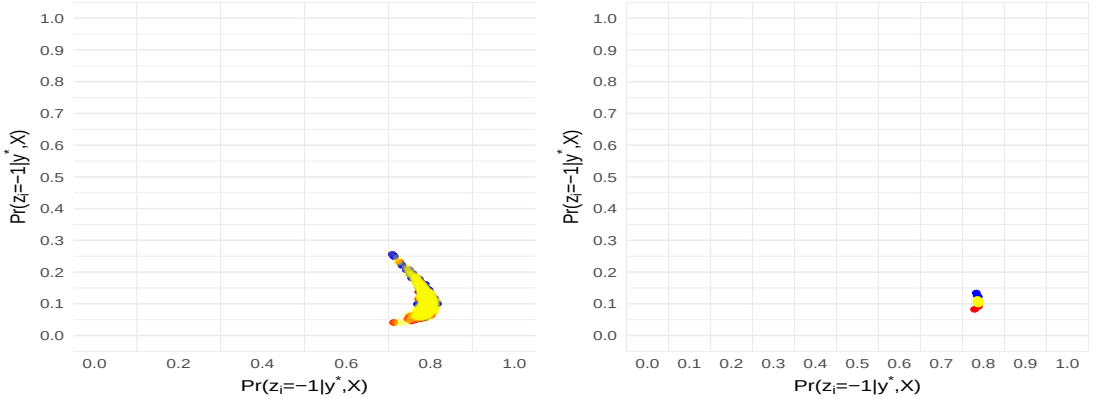
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

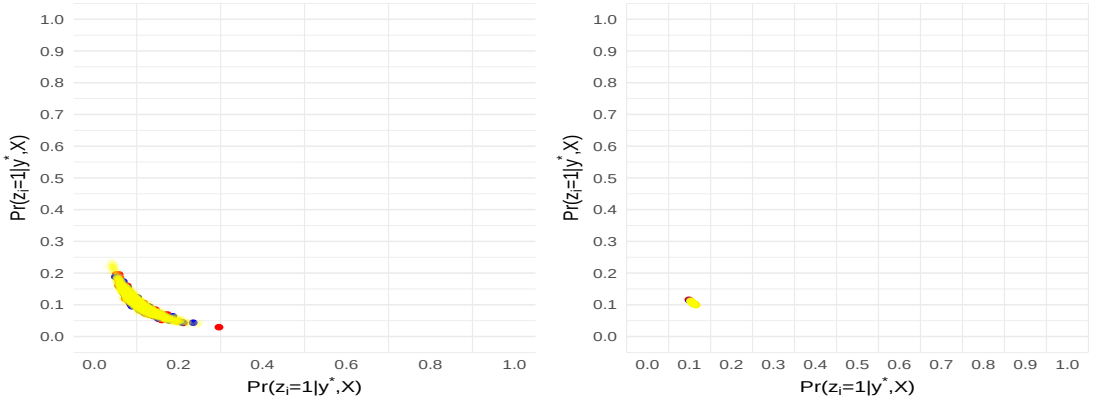
(d) averaged across simulations



(e) all simulation points

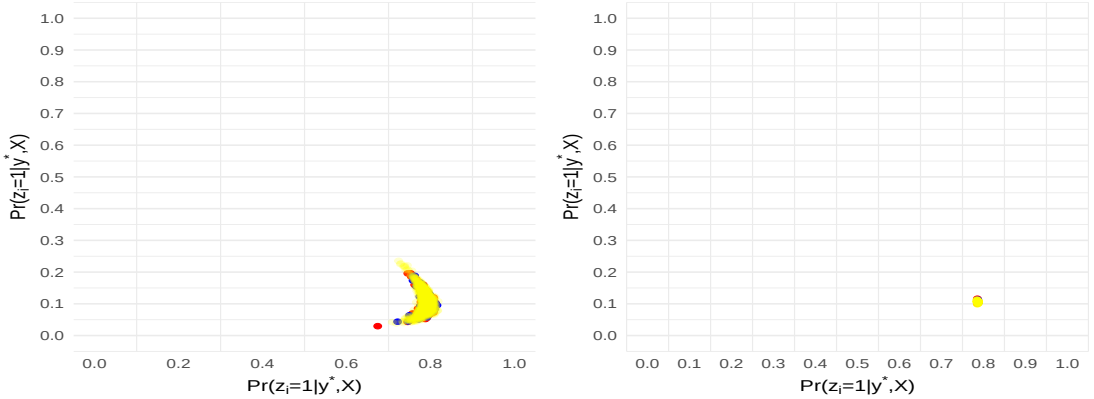
(f) averaged across simulations

Figure 3.7: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0.1$ under sparse prior for $\boldsymbol{z}$. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).



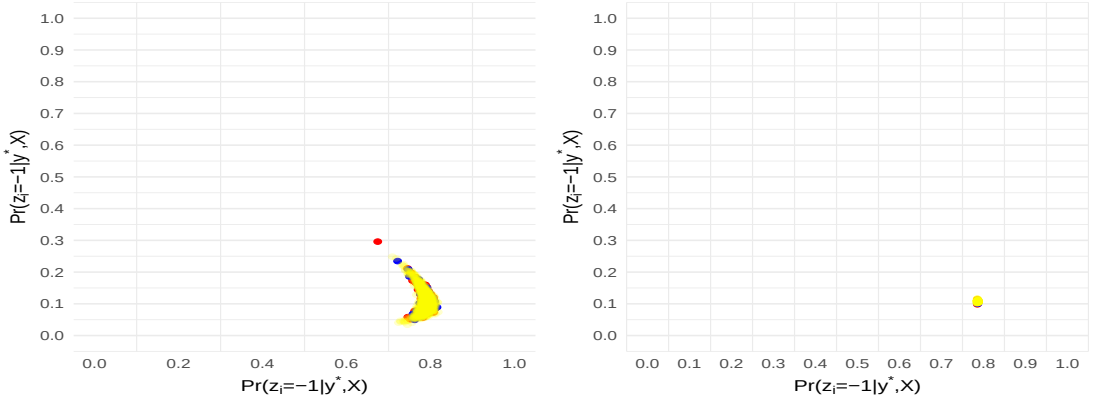
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

(d) averaged across simulations



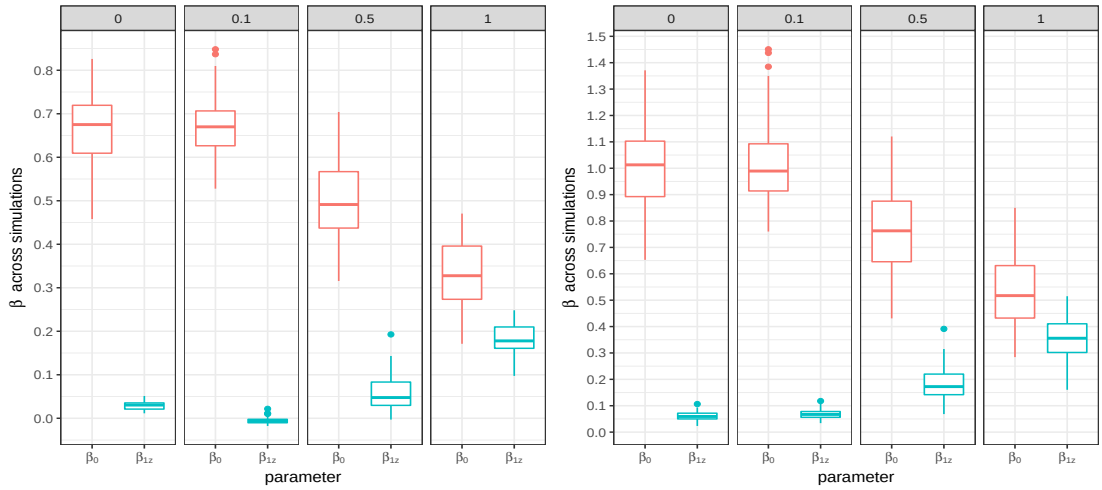
(e) all simulation points

(f) averaged across simulations

Figure 3.8: Simulation results for balance probit regression model with balance effect $\beta_{1z} = 0$ under sparse prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow).

3.3.3. *Posterior distribution of* β_{1z}

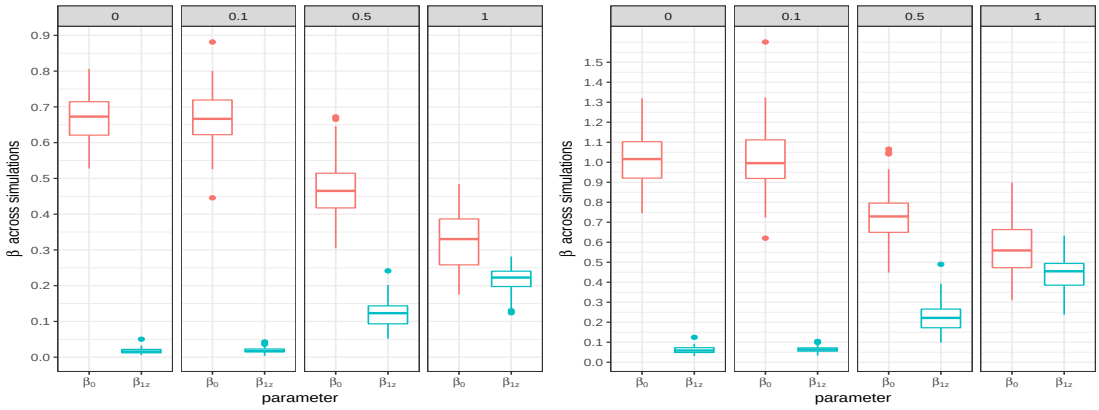
Figure 3.9 and Figure 3.10 show the posterior mean of β_{1z} over 100 simulations under uniform and sparse prior, respectively. Compared to Figure 2.5 and Figure 2.10 in Bayesian linear balance regression, the posterior estimates of β_{1z} tends to shrink to null, resulting in biased estimates from the posterior means. This is largely due to the randomness introduced during sampling of the unobserved latent vector $\mathbf{y}^*$. The posterior mean based on sampling $\beta_{1z}|\mathbf{y}^*, \boldsymbol{z}$ suffers from a larger shrinkage toward the null than that from $\beta_{1z}|\mathbf{y}, \boldsymbol{z}$ under both priors of $\boldsymbol{z}$.



(a) Conditional on $\mathbf{y}^*, \boldsymbol{z}$

(b) Conditional on $\mathbf{y}, \boldsymbol{z}$

Figure 3.9: Posterior mean of β_{1z} over 100 simulations in Bayesian probit balance regression under the uniform prior for $\boldsymbol{z}$.



(a) Conditional on $\boldsymbol{y}^*, \boldsymbol{z}$

(b) Conditional on $\boldsymbol{y}, \boldsymbol{z}$

Figure 3.10: Posterior mean of β_{1z} over 100 simulations in Bayesian probit balance regression under the sparse prior for $\boldsymbol{z}$.

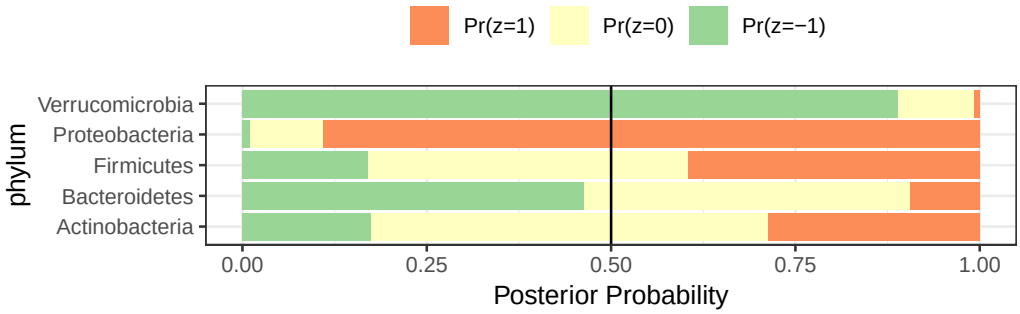
3.4. Application to real data analysis

Crohn's disease is one type of inflammatory bowel disease (IBD). Lewis et al., 2015 at University of Pennsylvania performed a study to examine the effect of inflammation, diet and antibiotic on the composition of gut microbiome. For this study, stool samples from 85 pediatric patients with Crohn's disease and 26 healthy volunteers were sequenced using shotgun metagenomic sequencing prior to any treatment or intervention. After quality control and filtering of rare taxa, 39 genera were obtained and their relative abundances were quantified. Low abundance genera with 0 values in more than 80% of all subjects were removed, resulting in 31 genera for our analysis. Zero values were replaced with half of the minimum abundance observed, a common practice in microbiome studies (Cao, Lin, and Li, 2017; Kurtz et al., 2015) and the proportions were recalculated after such zero replacements. We refer to this dataset as the IBD data set.

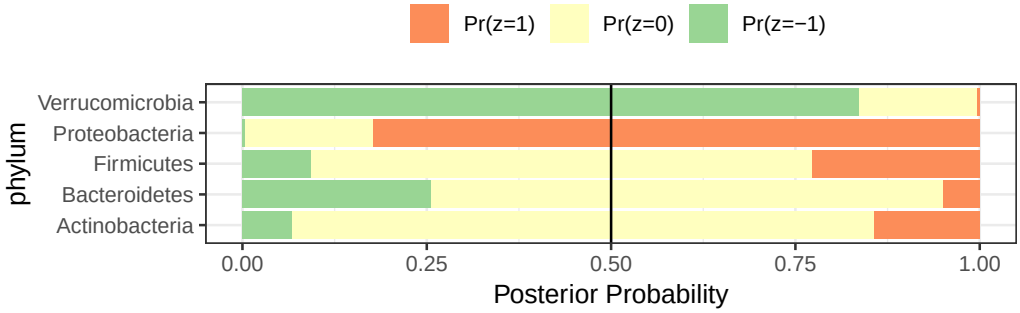
We use the same hyperparameter values as in Section 3.3 to perform proposed MCMC algorithm for the IBD data set. Five different starting values are randomly selected to check the convergence of Markov chains. Total number of iteration is 3×10^5 and burn-in is taken as the first 10^5 iterations. We also perform a sensitivity analysis with a non-informative and a sparse prior for the model space. Similar to Chapter 2, we also aggregate the data into two taxonomic levels: phylum level and genus level.

3.4.1. Analyses of the IBD data at the phylum level

With phylum level data, the total number of iterations is set to 10^5 and the burn-in step is 10^4. The posterior probabilities of each phylum being in the z_+, z_-, z_0 sets differ a little bit with two different priors, where the posterior probability being in z_+ or z_- for each phylum is lower under the sparse prior (see Table 3.1 and Table 3.2). Using 0.5 as the threshold for inference, the two priors agree on the bacteria that define the balance index, which is composed of log ratio of the relative abundance between *ProteoBacteria* and *Verrucomicrobia* (Figure 3.11).



(a) Uniform Prior



(b) Sparse Prior

Figure 3.11: Analysis of IBD data set at the phylum level. Posterior probability for 5 bacteria phyla being in the z_+, z_-, z_0 sets are shown. Top plot (a): uniform prior; bottom plot (b): sparse prior.

Table 3.1: Posterior probabilities with 5 starting points. Results are for genus level IBD data with uniform prior for z.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Bacteroidetes	0.09	0.46	0.44	0.09	0.46	0.45	0.10	0.47	0.44	0.10	0.46	0.44	0.10	0.46	0.44
Firmicutes	0.39	0.17	0.43	0.39	0.17	0.44	0.40	0.17	0.43	0.38	0.18	0.44	0.40	0.18	0.43

Table 3.1: Posterior probabilities with 5 starting points. Results are for genus level IBD data with uniform prior for z.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Verrucomicrobia	0.01	0.89	0.10	0.01	0.89	0.10	0.01	0.89	0.10	0.01	0.89	0.10	0.01	0.89	0.11
Proteobacteria	0.89	0.01	0.10	0.88	0.01	0.11	0.89	0.01	0.10	0.89	0.01	0.10	0.88	0.01	0.11
Actinobacteria	0.29	0.17	0.54	0.28	0.17	0.55	0.29	0.18	0.53	0.28	0.17	0.55	0.28	0.18	0.54

Table 3.2: Posterior probabilities with 5 starting points. Results are for genus level IBD data with sparse prior for z.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Bacteroidetes	0.05	0.26	0.69	0.04	0.25	0.71	0.04	0.25	0.71	0.05	0.26	0.70	0.04	0.24	0.72
Firmicutes	0.23	0.09	0.68	0.22	0.09	0.69	0.23	0.08	0.69	0.22	0.09	0.69	0.23	0.08	0.69
Verrucomicrobia	0.00	0.84	0.16	0.00	0.83	0.17	0.00	0.84	0.16	0.00	0.84	0.16	0.00	0.84	0.16
Proteobacteria	0.82	0.00	0.17	0.82	0.00	0.18	0.82	0.00	0.18	0.83	0.00	0.17	0.81	0.00	0.18
Actinobacteria	0.14	0.07	0.79	0.14	0.07	0.79	0.14	0.06	0.80	0.14	0.07	0.79	0.14	0.07	0.79

Summary of the posterior distributions of β under both the uniform and sparse priors is shown in Table 3.3. The posterior distribution of β is not sensitive to the choice of model space priors, both showing that the balance identified is associated with the risk of IBD.

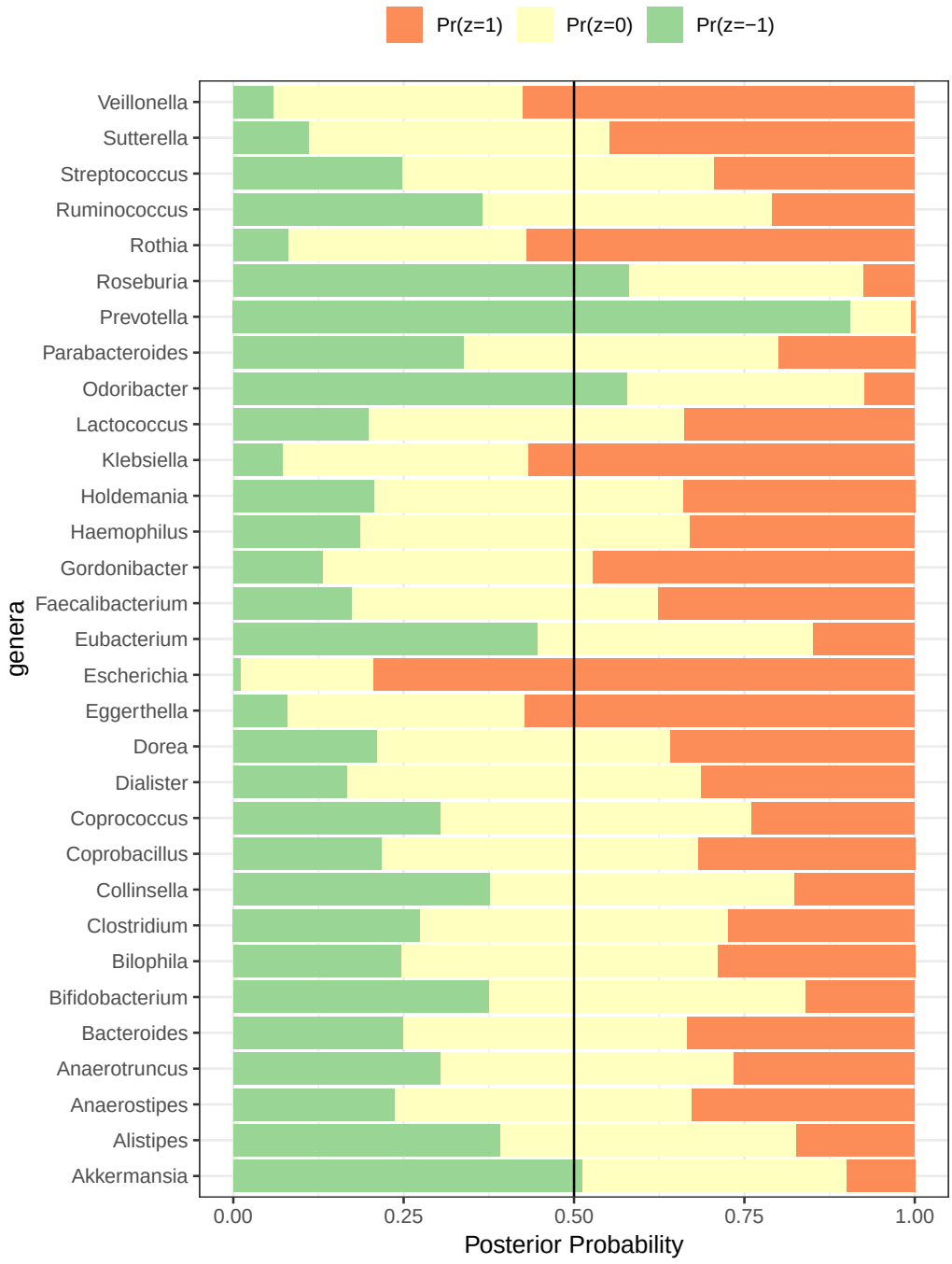
Table 3.3: Posterior distribution of β with phylum level data in IBD study. The posterior mean is calculated conditioning on y, z

	Conditional mean based on y, z			
	Non-informative prior		Sparse prior	
	β_0	β_1	β_0	β_1
Mean(sd)	0.32(0.41)	0.16(0.04)	0.29(0.43)	0.15(0.03)
95% credible interval	(-0.24,1.20)	(0.08,0.21)	(-0.26,1.20)	(0.08,0.20)

3.4.2. Analysis of the IBD data at genus level

Analysis using the genus level data is computed with 3×10^5 iteration and 5×10^4 burn-in steps. Analysis results are shown in Figure 3.12,Figure 3.13, Table A.5,Table A.6 and Figure 3.14.

The posterior probabilities of each genus being in the z_+, z_-, z_0 sets differ for two different priors. This is largely due to small sample size and a large number of bacterial taxa in the model. For the uniform prior, where we believe the each taxon is equally likely to be in one of the three sets (Figure 3.12), using 0.5 as the threshold for posterior inference, we conclude from Figure 3.12 that *Veillonella (Firmicutes)*, *Rothia (Actinobacteria)*, *Klebsiella (Proteobacteria)*, *Escherichia (Proteobacteria)*, *Eggerthella (Actinobacteria)* belong to the z_+ set and *Roseburia (Firmicutes)*, *Prevotella (Bacteroidetes)*, *Odoribacter (Bacteroidetes)*, *Akkermansia (Verrucomicrobia)* belong to the z_- set. As a comparison, using the sparse prior (Figure 3.13), *Escherichia* (*Proteobacteria*) is in the z_+ set and *Prevotella* (*Bacteroidetes*) in the z_- set.



(a) Uniform Prior

Figure 3.12: Posterior probabilities for 31 bacterial genera being in the z_+, z_-, z_0 sets under the uniform prior assumption.

(a) Sparse Prior

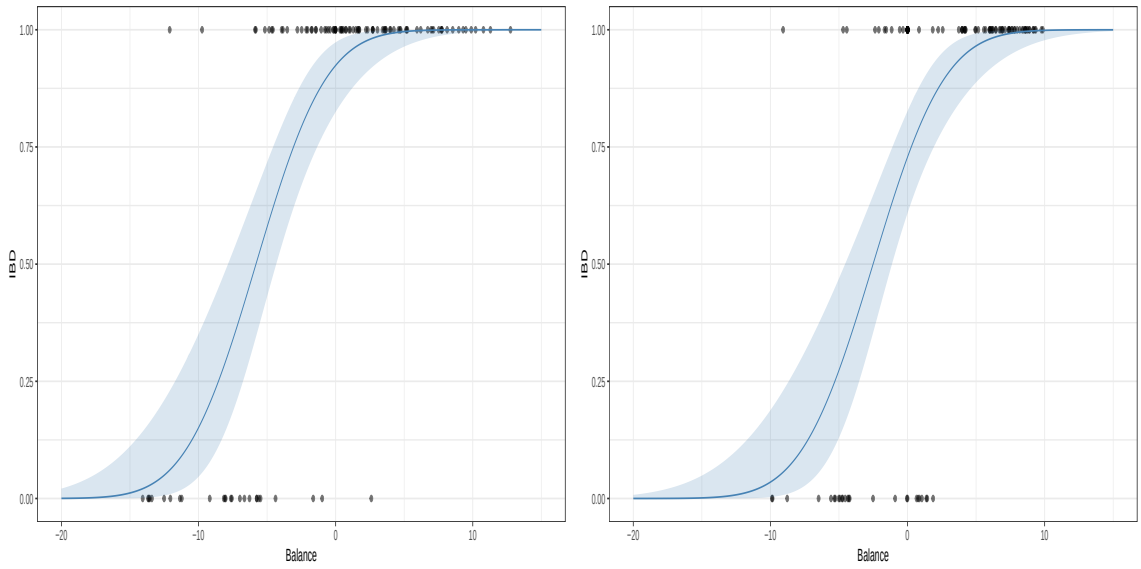
Figure 3.13: Posterior probabilities for 31 bacterial genera being in the z_+, z_-, z_0 sets under the sparse prior assumption.

Table 3.4: Posterior distribution of β for the IBD data at the genus level. The posterior mean is calculated conditioning on $\boldsymbol{y}, \boldsymbol{z}$.

	Conditional mean based on $\boldsymbol{y}, \boldsymbol{z}$			
	Non-informative prior		Sparse prior	
	β_0	β_1	β_0	β_1
Mean(sd)	1.43(0.70)	0.26(0.06)	1.16(0.61)	0.26(0.06)
95% credible interval	(0.22,2.95)	(0.15,0.38)	(0.18,2.61)	(0.16,0.38)

Table 3.4 shows a summary of the posterior distribution of the balance effect, indicating the association between the identified balance and the risk of IBD. The balance identified and our results are in general consistent with other finding in literature. We show that microbiota dysbiosis in IBD is characterized by an increase in *Proteobacteria* and a decrease in *Firmicutes* Lewis et al., 2015. In addition, our results also suggest other bacteria, especially *Actinobacteria, Bacteroidetes* and *Verrucomicrobia* might also be involved in the dysbiosis related to IBD.

Figure 3.14 shows the relationship between the IBD status and the estimated balance for models with uniform prior and sparse prior. In general, we observe that the estimated balance indeed differentiates IBD cases from the normal controls.



(a) Uniform Prior

(b) Sparse Prior

Figure 3.14: Scatter plot of IBD status vs the estimated balance with using genus level data in IBD study using (a) uniform prior and (b) sparse prior for $\mathbf{z}$. Solid line is estimated line from generalized linear model with probit link.

3.5. Discussion

The proposed Bayesian probit balance regression provides a method to discover the association between balance, a scalar that summarizes the microbiome compositional data, and the disease status, accounting for the potential interactions within microbiota. In this work, we extend the Bayesian linear balance regression in order to find the balance indicator that is associated with a binary outcome. Using the probit link with an auxiliary latent response variable facilitates the posterior sampling of the parameters, since the posterior and conditional distributions can be derived with closed-form expressions. Under this model, the data generation process can be partitioned into two steps. First, the balance index contributes to a continuous auxiliary response variable through a link function, and then the sign of the latent outcome determines the observed binary value. An efficient MCMC sampling algorithm is proposed and implemented. Numerical studies have proved its usefulness and illustrated the model performance. We have also applied the model to a real data set and identified a simple bacterial balance that is associated IBD.

In Bayesian probit balance regression, since there are no conjugate priors for the regression coeffi-

cients, the computation is more challenging and time-consuming. The probit link we used is a more natural extension of continuous outcome regression to binary outcome. Due to the extra sampling step for the auxiliary latent variable $\mathbf{y}^*$, it is more difficult to obtain accurate posterior inference in regression parameters for our proposed models. Compared to the Bayesian balance linear model with the same true parameter values, the posterior probabilities of the balance indicator in each of the z_+, z_-, z_0 tend to shrink more toward the prior means that are equal among all three sets. For our simulations, the truncated multivariate normal distribution for the latent variable $\mathbf{y}^*$ has a dimensionality of $n = 100$, which can be very time consuming to sample. In our implementation, a Gibbs sampling is used to sample one coordinate at a time. With such a high dimension, it took a long time to find a single point in the truncated space, and even longer time to find a good point in such a truncated space. We have tried different sample sizes, and $n = 100$ is the maximum sample size that we can manage to complete the simulation studies.

Alternatively, one can develop Bayesian logistic balance regression, in which case Laplacian approximation of the posterior can be applied. Such an approximation does not limit what link function to use in the regression, but it might lead to reduced accuracy because it relies on a linearization of the posterior function. After approximation, a multivariate normal priors can be imposed to obtain a closed-form posterior distribution of the model. An future research topic is to use the Laplacian approximation and the logistic link for logistic balance regression analysis.

CHAPTER 4

BAYESIAN BALANCE MEDIATION ANALYSIS

4.1. Introduction

Microbiome, the totality of all the microbes on and in human body, has been shown to be associated with human health and disease. Those microbes exert their functions in a coordinated fashion and their abundance relative to each other is relatively stable from day to day. However, dysbiosis, defined as large fluctuations in the relative abundance of the microbes can lead to disease. For example, *Clostridium difficile* infection is a potentially lethal disease in the large intestine that is caused by the toxic spores from excessive proliferation of *Clostridium difficile* bacteria. Although *Clostridium difficile* is reported to reside in healthy people, antibiotic treatment can lead to disturbance in gut microbiota, which can lead to increase of *Clostridium difficile* in the gut and the corresponding clinical conditions. In this case, gut microbiome serves as the mediator of the effect of antibiotic use on clinical outcomes. Recent study has shown that intake of metformin changes the gut microbiota composition in normoglycaemic young, which may be a determinant for development of gastrointestinal adverse effects following metformin intake (Bryrup et al., 2019). In this study, gut microbiome plays the mediating role to link metaformin treatment to gastrointestinal adverse effects.

Mediation analysis can be used to address important scientific questions in many microbiome studies, where the goal is to understand the role of gut microbiome in linking the effects of treatment or risk factors on the outcome. Classical mediation analysis usually considers a single mediator or intervention variable and has strong theoretical foundations and extensive applications in medical research and in economics (Imai, Keele, and Yamamoto, 2010; Pearl, 2000; Rubin, 2005). With a continuous outcome, mediation analysis is often performed through structural linear equation model. shown in (4.1), where M is the mediator, T represents treatment variable and Y is the outcome variable. The mediation effect (or indirect effect) of T through M is the product of two two

path coefficients a and b and the direct effect of T is the path coefficient c.

$$E(M) = a_0 + aT, \tag{4.1}$$
$$E(Y) = b_0 + cT + bM.$$

The path coefficients can be estimated based on the two linear models and the standard errors for indirect effect ab is usually calculated by multivariate delta method, which relies on the assumption that the asymptotic distribution of ab is approximately normal (Sobel, 1982). However, the distribution of ab is usually skewed and several improvements have been proposed to address such an asymmetry by constructing confidence intervals using the bootstrap method or using the product of two normally distributed random variables (Bollen and Stine, 1990; Cheung, 2009; Shrout and Bolger, 2002).

However, mediation analysis with microbiome as possible mediator is challenging. In typical 16S rRNA sequencing studies, the microbiome data can be summarized as a matrix of sequencing read counts, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$, for n samples and q taxa, where $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iq})$ is the vector of read counts of q taxa in the ith sample. Such read counts are usually normalized into the matrix of relative abundances of the bacterial taxa or are assumed to be sampled from a multinomial distribution with the underlying composition matrix $\boldsymbol{P} = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n)^T$. Modeling such compositional data is challenging due to the unit sum constraint (Li, 2015). Existing work extends the single mediator framework with uncorrelated mediators or transformation of correlated mediators (Chén et al., 2017; VanderWeele and Vansteelandt, 2014). The corresponding path coefficient a is estimated one mediator at a time. However, those methods can not be applied to microbiome data due its high-dimensionality and unit-sum constraint of the bacterial composition. Sohn and Li, 2019 developed the first methodology for compositional mediation analysis. In this work, the path coefficient vector is jointly estimated without distributional assumptions and the high-dimensional problem is alleviated through regularization. The joint and component-wise inference on the mediation effect can be tested with properly defined null hypothesis and bootstrap confidence intervals.

Log-ratio of partitions of a composition $\boldsymbol{p} = (p_1, \cdots, p_q)$, termed balance, was first introduced in geology by Egozcue and Pawlowsky-Glahn (Egozcue and Pawlowsky-Glahn, 2005), and was recently applied to microbiome studies with meaningful discoveries (Huang and Li, 2020; Morton et

al., 2017). Balance is defined as a weighted log ratio between two partitions of a compositional vector. Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_q)$ be a q-dimensional indicator vector with three values, 1, -1, and 0, where $z_i = 1$ indicates that the corresponding part in the composition p_i is in the numerator of a ratio, $z_i = -1$ indicates the denominator; and $z_i = 0$ indicates the parts that are not a member of the balance. Let z_+, z_-, and z_0 be non-overlapping sets of indices of the elements in $\boldsymbol{z}$ whose values are 1, -1 and 0, respectively, and $z_+ \cup z_- \cup z_0 = \{1, 2, \ldots, q\}$. Given $\boldsymbol{z}$, the balance is defined as

$$B_z = \sqrt{\frac{1}{1/|z_+| + 1/|z_-|}} \log\left(\frac{\prod_{i \in z_+} p_i}{\prod_{j \in z_-} p_j}\right), \tag{4.2}$$

where $|\cdot|$ denotes the cardinality of a set. Balance roughly captures relative fold change between two subgroups of taxa that compete against each other where the microbes within each subgroup have symbiotic relationship. It provides a simple and useful quantity in describing the microbiota community (Rivera-Pinto et al., 2018). In previous applications, the value of the indicator vector $\boldsymbol{z}$ is chosen *a priori* based on a phylogenetic tree or some other biological knowledge (Morton et al., 2017; Washburne et al., 2017). However, such information may be biased or even misleading since the taxonomic positions do not always differentiate beneficial from detrimental taxa.

Based on this concept of balance, we propose a Bayesian balance mediation analysis, where balance serves as the mediator. Our model assumes that the treatment leads to change of balance of the microbial community, which leads to the observed outcome. By introducing the balance as the single mediator, we simplify the estimation of path coefficient b for the effect of treatment transmitted through microbiota. The challenge of our proposed Bayesian balance mediation analysis is to identify the sets z_+ and z_- that define the balance. We use the compositional algebra to quantify how treatment shifts microbiome composition and the corresponding balance, where the path coefficient vector $\boldsymbol{a}$ is also compositional. We show that the indirect mediation effect of the treatment is the multiplication of two quantities: the path coefficient b and the effect of balance determined by the path vector $\boldsymbol{a}$ and balance indicator vector $\boldsymbol{z}$. We then develop a MCMC algorithm to identify the sets z_+ and z_- and to obtain the posterior predictive distribution of the mediation effect through the estimated balance.

The main contribution of this paper is to incorporate the balance index into the compositional mediation analysis. In such situation, the path coefficient a in Equation (4.1) will be replaced by a vector value, represented by $\boldsymbol{a}$. By introducing the balance index calculated from a compositional vector

as the single mediator, we simplify the estimation of path coefficient b for the effect of treatment that is transmitted through microbiota into a scalar value. The path coefficient vector $\boldsymbol{a}$ of how treatment affects the microbiota is modeled as a hierarchical Bayesian model conditional on a balance index. We also proved that mediation effect of treatment is a multiplication of two quantities given the balance indicator $\boldsymbol{z}$: the path coefficient b and applying the same transformation as the balance index to the path cofficient vector a. We designed a sampling approach to sample parameters in the proposed Bayesian Balance Mediation Model where the treatment is dichotomous and provided posterior inferences on parameters of interest. Our method is different compared to Sohn and Li, 2019 in three main aspects: we assume a full parametric Bayesian model; the counts are realization of underlying distribution; we can not estimate component wise mediation effect for a particular taxon.

4.2. Bayesian balance mediation analysis

4.2.1. Notation

Consider a microbiome study with n $i.i.d.$ samples from a population. Let $\boldsymbol{y} = (y_1, \cdots, y_n)$ be the vector of a continuous outcome of length n, and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iq})$ be the observed read counts of q bacterial taxa with a total number of counts n_i. The treatment assignment for subject i is represented by a variable t_i. For binary treatment, it takes value of 0 for sample that is untreated and 1 for treated. Let $\boldsymbol{p}_i = (p_{i1}, \cdots, p_{iq})$ be a q-dimensional compositional vector representing the relative abundance of q bacterial taxa for the ith sample. Let $\boldsymbol{z}$ is the q-dimensional indicator vector for balance, where $z_j = 1$ indicates p_{ij} in the z_+ of the balance index, and $z_j = -1$ indicates p_{ij} in the z_- of the balance index, and $z_j = 0$ indicates that p_{ij} does not contribute to balance. We further let $bal()$ represent the balance defined by the compositional vector $\boldsymbol{p}_i$ and the balance indicator vector $\boldsymbol{z}$, and let $B_{zi} = bal(\boldsymbol{z}, \boldsymbol{p}_i)$ be the balance calculated from $\boldsymbol{p}_i$ and $\boldsymbol{z}$ for the ith sample.

Before introducing the Bayesian Balance Mediation Model (BBMM), we define several compositional operators. The additive logratio transformation $\phi(\boldsymbol{p_i})$ for a compositional vector $\boldsymbol{p_i}$ is defined as follows:

$$\phi(\boldsymbol{p_i}) = \left(\log \frac{p_{i1}}{p_{iq}}, \log \frac{p_{i2}}{p_{iq}}, \ldots, \log \frac{p_{i(q-1)}}{p_{iq}}\right). \tag{4.3}$$

Since $\phi(\cdot)$ is a one-to-one transformation, the inverse of additive logratio transformation $\phi^{-1}(\cdot)$ is

uniquely defined. A $\oplus$ operator of two compositional vectors $\boldsymbol{p_i}$ and $\boldsymbol{p_j}$ is equivalent to renormalized component-wise multiplication (Aitchison, 1982)

$$\boldsymbol{p_i} \oplus \boldsymbol{p_j} = \left(\frac{p_{i1}p_{j1}}{\sum_{k=1}^{q} p_{ik}p_{jk}}, \frac{p_{i2}p_{j2}}{\sum_{k=1}^{q} p_{ik}p_{jk}}, \ldots, \frac{p_{iq}p_{jq}}{\sum_{k=1}^{q} p_{ik}p_{jk}} \right). \tag{4.4}$$

Similarly, the power operator with a compositional base $\boldsymbol{a}$ and a scalar in the exponent t is defined as

$$\boldsymbol{p}^t = \left(\frac{p_1^t}{\sum_{j=1}^{q} p_j^t}, \frac{p_2^t}{\sum_{j=1}^{q} p_j^t}, \ldots, \frac{p_q^t}{\sum_{j=1}^{q} p_j^t} \right). \tag{4.5}$$

4.2.2. Bayesian Balance Mediation Model

The key of our proposed BBMM is to associate the treatment t_i, the observed count vector $\boldsymbol{x}_i$ and the outcome y_i via the unobserved bacterial relative abundance vector $\boldsymbol{p}_i$ and the balance configuration vector $\boldsymbol{z}$. We propose the following general hierarchical model:

$$\boldsymbol{x_i} \sim \text{ Multinomial}(n_i, \boldsymbol{p_i}) \tag{4.6}$$

$$\boldsymbol{p_i} = \boldsymbol{m_0} \oplus \boldsymbol{a}^{t_i} \oplus \boldsymbol{U_i} \tag{4.7}$$

$$Y_i | \boldsymbol{p_i}, \boldsymbol{z} = a_z + c_z \cdot t_i + b_z \cdot B_{zi} + e_{zi} \tag{4.8}$$

where $\boldsymbol{m_0}$ is the baseline composition, $\boldsymbol{a}$ is the compositional parameter that measures the effect of t_i on composition $\boldsymbol{p}_i$, and U_i is the random error for a compositional vector. We further let $\boldsymbol{\beta}_z = (a_z, c_z, b_z)$, $\boldsymbol{w}_i = (1, t_i, B_{zi})$, and $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)^T$.

For simplicity of presentation, we consider the case when the treatment t_i takes binary 0/1 values, as often seen in randomized clinical trial setting. We reparameterize the BBMM using mean values $\boldsymbol{\mu_0}, \boldsymbol{\mu_1} \in R^{q-1}$ for the compositions after the additive logratio transformation. In addition, we propose to use a logistic normal distribution (Billheimer, Guttorp, and Fagan, 2001) with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma} \in R^{(q-1)*(q-1)}$ to characterize the random noise vector $\boldsymbol{U_i}$. This implies that

$$\phi(\boldsymbol{p}_i) = \left(\log \frac{p_{i1}}{p_{iq}}, \log \frac{p_{i2}}{p_{iq}}, \log \frac{p_{i(q-1)}}{p_{iq}} \right) \sim MVN(\boldsymbol{\mu_{t_i}}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\mu}_{t_i} = \boldsymbol{\mu}_0(1-t_i) + \boldsymbol{\mu}_1 t_i$, we then have an equivalent model for $\boldsymbol{p}_i$ as model (4.7):

$$\boldsymbol{p}_i \sim \text{logistic normal}(\boldsymbol{\mu}_0(1-t_i) + \boldsymbol{\mu}_1 t_i, \boldsymbol{\Sigma}).$$

By imposing distributional assumptions on $\boldsymbol{U_i}$ and e_{zi} and prior distributions on parameters in the BBMM, we can then draw posterior inferences on all model parameters. To complete the model, we assume the standard conjugate priors for a_z, c_z, b_z, μ_0, μ_1, σ^2 and $\boldsymbol{\Sigma}$. Specifically, we assume

$$\begin{aligned}
&\boldsymbol{\beta}_z = (a_z, c_z, b_z) \sim MVN(\boldsymbol{\beta}_0, \boldsymbol{V}), \quad (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)^T \sim MVN(\boldsymbol{\eta}, \boldsymbol{\Omega}), \\
&\sigma^2 \sim invGamma(\nu/2, \lambda/2), \quad \boldsymbol{\Sigma}^{-1} \sim W_{q-1}(\rho, \boldsymbol{\Psi}^{-1}),
\end{aligned}$$

where $\beta_0, \boldsymbol{V}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \nu, \lambda, \rho$ and $\boldsymbol{\Psi}$ are the hyper-parameters. In this paper, following Billheimer, Guttorp, and Fagan, 2001, the hyperparameters are chosen as the following:

$$\begin{aligned}
\boldsymbol{\eta} &= \boldsymbol{0_{q-1}}, \\
\boldsymbol{\Omega} &= k_1(\boldsymbol{I_{q-1}} + \boldsymbol{j_{q-1}} \boldsymbol{j}_{q-1}^T),
\end{aligned}$$

where k_1 is a hyper parameter that we need to choose for the prior distribution and is fixed at 0.5, $\boldsymbol{I_{q-1}}$ is the identify matrix, and $\boldsymbol{j}_{q-1}$ is the vector of ones. This corresponds to an evenly distributed compositional vector whose 95% probability contour is around 0.05 for each component. The degree of freedom ρ is often chosen as $q-1$, and $\boldsymbol{\Psi}$ is a positive definite matrix, often chosen as:

$$\boldsymbol{\Psi} = k_2(\boldsymbol{I_{q-1}} + \boldsymbol{j_{q-1}} \boldsymbol{j}_{q-1}^T)$$

where k_2 is chosen as 0.1. These corresponds to a weak prior for $\boldsymbol{\Sigma}^{-1}$.

Finally, a straightforward choice for the prior of $\boldsymbol{z}$ is independent multinomial distribution for each component of $\boldsymbol{z}$.

$$\pi(\boldsymbol{z}) = w_1^{m_+} w_2^{m_-} (1 - w_1 - w_2)^{p - m_+ - m_-} \tag{4.9}$$

where w_1 and w_2 are the expected number of variables in z_+ and z_- group respectively. They affect the sizes of taxa with positive and negative effects in the posterior inference of balance. A non-informative choice is $w_1 = w_2 = 1/3$, resulting in $\pi(\boldsymbol{z}) = (1/3)^p$. On the other hand, small values

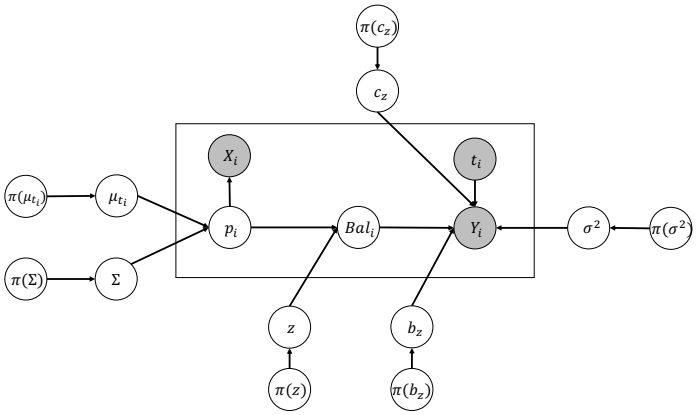


Figure 4.1: A graph diagram of the proposed Bayesian balance mediation analysis, where the shaded circles present the data observed.

of w_1 and w_2 assume sparse structure of the balance. This prior has implicit assumptions that two structures of a balance is as good as each other when the size of z_+ and z_- are approximately equal. It is possible to relax the independence specification and use a dependent prior among elements of $\boldsymbol{z}$. Figure 4.1 shows the summary and the conditional independence structure of the proposed BBMM.

4.2.3. Direct and mediation effect

In classical mediation analysis model (Equation (4.1)), under certain assumptions, the direct effect c and mediation effect ab of treatment on the continuous outcome can be estimated from the coefficients in two linear models. The standard assumptions in mediation analysis include $0 < P(t_i = t) < 1$, $0 < P(\boldsymbol{p_i} = \boldsymbol{p}|t_i = t) < 1$, no interference among the subjects and no interactions between t_i and $\boldsymbol{p}_i$. Besides these standard assumptions, two key conditional independence assumptions under the potential outcome framework are also needed:

$$\{y_i(t', \log p), \boldsymbol{p_i}(t)\} \perp\!\!\!\perp t_i \tag{4.10}$$

$$y_i(t', \log p) \perp\!\!\!\perp \boldsymbol{p_i}(t)|t_i \tag{4.11}$$

where $y_i(t', \log p)$ is the potential outcome when the treatmenet is set to t' and the log-composition is set to p, and $\boldsymbol{p_i}(t)$ is the bacterial composition when the treatment is set to t. Assumption (4.10)

requires that the treatment assignment is independent of the outcome and the relative abundance vector $\boldsymbol{p}_i$, and assumption (4.11) states that given treatment, outcome is independent of the relative abundance vector $\boldsymbol{p}_i$. Under these assumptions, we have the following Lemma linking the model parameters to the direct and indirect mediation effects for a given balance configuration vector $\boldsymbol{z}$.

Lemma 1 *Under the above assumptions, for a given balance configuration vector $\boldsymbol{z}$ and the proposed BBMM, specified by equations* (4.6) - (4.8)*, the direct treatment effect on the outcome is c_z and the indirect mediation effect of the balance is the product of $bal(\boldsymbol{z}, \boldsymbol{a})$ and b_z. For binary treatment, the mediation effect of the balance is the product of $bal(\boldsymbol{z}, \phi^{-1}(\mu_1 - \mu_0))$ and b_z.*

4.3. Model fit and inference via MCMC sampling

4.3.1. MCMC sampling

To obtain the estimates for all the unknown parameters in the model illustrated in Figure 4.1, we propose a MCMC sampling procedure that combines the Gibbs and Metropolis-Hasting sampling. The key component linking the two equations in Model (4.6) is $\boldsymbol{p}_i$. Conditioning on $\boldsymbol{p}_i$, we can conduct separate estimation for the two equations. In brief, the MCMC steps can be summarized as

$$\begin{aligned}\boldsymbol{\Sigma}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{p}^{(0)}, \boldsymbol{z}^{(0)} \Rightarrow & \begin{cases} \text{sample } \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)} \text{ based on } \boldsymbol{X} \\ \text{sample } \boldsymbol{z}^{(1)} \text{ based on } \boldsymbol{y} \end{cases} \\ & \Rightarrow \text{sample } \boldsymbol{p}^{(1)} \text{ based on } \boldsymbol{X}, \boldsymbol{y} \\ & \Rightarrow \text{sample } b^{(1)}, c^{(1)}, \sigma^{2(1)} \ldots\end{aligned}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)^T$. The detailed sampling steps for fitting the Bayesian balance mediation models is presented as Algorithm 3.

Algorithm 3: MCMC algorithm for the Bayesian balance mediation analysis

a. Initialize $\boldsymbol{\Sigma}^{(0)}, \boldsymbol{\mu_0}^{(0)}, \boldsymbol{\mu_1}^{(0)}$ by sampling from their prior distributions.
b. Initialize $\boldsymbol{p}_i^{(0)}$ for each individual by sampling from $f(\boldsymbol{p}_i|\boldsymbol{\mu}_{\boldsymbol{t}_i}, \boldsymbol{\Sigma})$.
c. Initialize $\boldsymbol{z}$ randomly or from the prior distribution.
d. Perform the following sampling iteratively, for the kth iteration:
   (a) Sample $\boldsymbol{\Sigma}^{(k)}, \boldsymbol{\mu_0}^{(k)}, \boldsymbol{\mu_1}^{(k)}$ from conditional distribution with closed form

$$f(\boldsymbol{\Sigma}^{(k)}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{p_i^{(k-1)}}, \boldsymbol{\mu_0^{(k-1)}}, \boldsymbol{\mu_1^{(k-1)}}),$$
$$f(\boldsymbol{\mu_0}^{(k)}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{p_i^{(k-1)}}\boldsymbol{\Sigma}^{(k)}),$$
$$f(\boldsymbol{\mu_1}^{(k)}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{p_i^{(k-1)}}, \boldsymbol{\Sigma}^{(k)}).$$

   (b) As given $\boldsymbol{p}_i$, X_i, y_i are independent, we can sample $\boldsymbol{z}^{(k)}$ with Hastings sampling with joint probability $f(\boldsymbol{z^{(k)}}, \boldsymbol{y}|\boldsymbol{p_i^{(k-1)}})$.
   (c) Sample $\boldsymbol{p_i^{(k)}}$ with Hasting sampling, with logistic normal distribution centered at current value from the joint probability $f(\boldsymbol{p_i^{(k)}}, \boldsymbol{y}, \boldsymbol{X}|\boldsymbol{z^{(k)}}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\mu_0}^{(k)}, \boldsymbol{\mu_1}^{(k)})$
   (d) Sample $\boldsymbol{\beta}_z, \sigma^2$ from the closed form conditional distributions with current value of $\boldsymbol{z}$ and all unknown values.

4.3.2. Conditional distributions

We present some details of the conditional distributions that are needed in the MCMC sampling presented in Algorithm 3. The joint distribution of ($\boldsymbol{X}$, $\boldsymbol{p}$, μ, $\boldsymbol{\Sigma}$) is

$$\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto & f(\boldsymbol{\Sigma}) \prod_i \{f(\boldsymbol{x_i}|\boldsymbol{p}_i) f(\boldsymbol{p}_i|\boldsymbol{\mu}_{\boldsymbol{t_i}}, \boldsymbol{\Sigma}) * f(\boldsymbol{\mu}_{\boldsymbol{t_i}})\} \\
= & |\boldsymbol{\Psi}|^{\rho/2} |\boldsymbol{\Sigma}|^{-(\rho-q)/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\} \\
& \times \prod_{i=1}^{n} \left(\prod_{j=1}^{q} p_{ij}^{x_{ij}-1} \right) \\
& \times \prod_{i=1}^{n} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\phi(\boldsymbol{p}_i) - \boldsymbol{\mu}_{\boldsymbol{t_i}})^T \boldsymbol{\Sigma}^{-1} (\phi(\boldsymbol{p}_i) - \boldsymbol{\mu}_{\boldsymbol{t_i}})\right\} \\
& \times \prod_{i=1}^{n} |\boldsymbol{\Omega}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{t_i}} - \boldsymbol{\eta})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{\mu}_{\boldsymbol{t_i}} - \boldsymbol{\eta})\right\}.
\end{aligned}$$

Basing on this, we can obtain the following conditional distributional distributions for $\boldsymbol{\mu}_k, k = 0, 1$ and $\boldsymbol{\Sigma}$ (Gelfand et al., 1990),

$$\begin{aligned}
\boldsymbol{\mu_k}|\boldsymbol{\Sigma}, \boldsymbol{p} &\sim MVN(\boldsymbol{\mu}_{\boldsymbol{k}}^*, \boldsymbol{\Sigma}_k^*), \text{ for } k = 0, 1, \\
\boldsymbol{\Sigma}^{-1}|\boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{P} &\sim W\left\{n + \rho - q, \left(\boldsymbol{\Psi} + \sum_i \left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu}_{t_i}\right]\left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu}_{t_i}\right]^T\right)^{-1}\right\}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{k}}^* &= (\boldsymbol{\Omega}^{-1} + n_k\boldsymbol{\Sigma}^{-1})^{-1}\left\{n_k\boldsymbol{\Sigma}^{-1}\frac{\sum_i \phi(p_i) * I(t_i = k)}{n_k} + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right\}, \\
\boldsymbol{\Sigma}_k^* &= (\boldsymbol{\Omega}^{-1} + n_k\boldsymbol{\Sigma}^{-1})^{-1}.
\end{aligned}$$

These conditional distributions are used to update $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$ and $\boldsymbol{\Sigma}$ in Step 4(a).

In order to sample $\boldsymbol{p}_i$ for subject i, we notice that $\boldsymbol{x}_i, y_i$ are independent given $\boldsymbol{p}_i$. Treating all the other unknowns as fixed, the conditional distribution of $\boldsymbol{p}_i$ given the balance configuration is

$$\begin{aligned}
& f(\boldsymbol{p}_i|\boldsymbol{x_i}, y_i, \boldsymbol{z}) \propto f(\boldsymbol{x_i}|\boldsymbol{p}_i) \times f(y_i|\boldsymbol{p_i}, \boldsymbol{z}) \times f(\boldsymbol{p}_i) \\
= & \left(\prod_{j=1}^q p_{ij}^{x_{ij}}\right) \times \int f(y_i|\boldsymbol{p_i}, \boldsymbol{z}, b_z, c_z, \sigma^2) db_z dc_z d\sigma^2 \\
& \times (\frac{1}{2\pi})^{(q-1)/2}|\boldsymbol{\Sigma}|^{-1/2}(\prod_{j=1}^q p_{ij})^{-1}\exp\left(-\frac{1}{2}\left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu_{t_i}}\right]^T\boldsymbol{\Sigma}^{-1}\left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu_{t_i}}\right]\right) \\
\propto & \left(\prod_{j=1}^q p_{ij}^{x_{ij}-1}\right) \times |\boldsymbol{V}^*|^{1/2}(v\lambda + \boldsymbol{\beta}_0^T\boldsymbol{V}^{-1}\boldsymbol{\beta}_0 + \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{u}^T(\boldsymbol{V}^*)^{-1}\boldsymbol{u})^{-(v+n)/2} \\
& \times \exp\left(-\frac{1}{2}\left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu_{t_i}}\right]^T\boldsymbol{\Sigma}^{-1}\left[\phi(\boldsymbol{p}_i) - \boldsymbol{\mu_{t_i}}\right]\right)
\end{aligned}$$

where v, λ are the parameters for prior distribution of σ^2 and

$$\boldsymbol{V}^* = (\boldsymbol{V}^{-1} + \boldsymbol{W}^T\boldsymbol{W})^{-1}, \tag{4.12}$$

$$\boldsymbol{u} = \boldsymbol{V}^*(\boldsymbol{V}^{-1}\boldsymbol{\beta}_0 + \boldsymbol{W}^T\boldsymbol{y}). \tag{4.13}$$

We propose to sample $\boldsymbol{p}_i$ from this distribution using the Metropolis-Hastings algorithm, where the proposal distribution is logistic -normal centered at current $\boldsymbol{p}_i$ and a fixed positive definite covariance

matrix.

We then sample the balance configuration $\boldsymbol{z}$ using a similar sampling scheme as in Huang and Li, 2020. For stochastic search of $\boldsymbol{z}$, three possible value changes can occur during the random exploration: i) between 0 and 1; ii) between 0 and -1; iii) between 1 and -1. Modifications between two values has three alterations: value conversions in either direction and value switch. We set equal probabilities among all possible modifications, but such probabilities can be adjusted to speed up the exploration of the sampling space. The joint distribution $\prod_i f(\boldsymbol{p}_i, \boldsymbol{x_i}, y_i, \boldsymbol{z})$ is used to determine on accepting or rejecting the newly proposed $\boldsymbol{z}$. Lastly, we flip the sign for $\boldsymbol{z}$ if the model fitting between outcome and all the other variables in the regression model results in a negative coefficient for balance. This is to ensure the identifiability of $\boldsymbol{z}$. Finally, after we sample $\boldsymbol{p}_i$ and $\boldsymbol{z}$, we sample $\boldsymbol{\beta}_z$ and σ^2 using the same sampling procedure as in Huang and Li, 2020.

4.3.3. Posterior inference

The convergence and mixing of the MCMC algorithm can be evaluated using different chains with different starting points and by examining the cumulative moving average for each component of $\boldsymbol{z}$. After convergence, the posterior marginal inclusion probabilities for sets z_+, z_-, z_0 are simply taken as the marginal proportions of $1, -1, 0$ in $\boldsymbol{z}$. A threshold can be set to make inference on which set of particular taxa should belong to the balance and the resulting balance definition.

Conditioning on $\boldsymbol{z}$, we can obtain the predictive posterior distribution of the regression coefficients $\boldsymbol{\beta}_z$. From model (4.8) and the normal prior distribution of $\boldsymbol{\beta}_z$, one can show that $\boldsymbol{\beta}_z|\boldsymbol{Y}, \boldsymbol{z}$ has a multivariate t-density with degrees of freedom $\lambda + n$, location parameter $\boldsymbol{u}$, and shape matrix

$$\tilde{\boldsymbol{V}} = \frac{v\lambda + \boldsymbol{b_0}^T \boldsymbol{V}^{-1} \boldsymbol{b_0} + \boldsymbol{Y}^T \boldsymbol{y} - \boldsymbol{u}^T \boldsymbol{V^*}^{-1} \boldsymbol{u}}{v + n} \boldsymbol{V^*}, \tag{4.14}$$

where $\boldsymbol{V^*}$ and $\boldsymbol{u}$ are defined in (4.12) and (4.13). During each iteration of MCMC, we sample $\boldsymbol{\beta}_z$ from its posterior distribution and perform posterior inference for $\boldsymbol{\beta}_z$ based on these sampling points. Finally, at the convergence, due to randomness of $\boldsymbol{z}$, the posterior samples of $\boldsymbol{\beta}_z$ are averaged over those $\boldsymbol{z}$, which gives the mean of the predictive posterior distribution of $\boldsymbol{\beta}_z$ based on model average. Similarly, we can obtain the posterior predictive distribution of the mediation effect $bal(\boldsymbol{z}, \boldsymbol{a}) \times b_z$.

4.4. Simulation studies

We evaluate the proposed Bayesian balance mediation analysis for binary treatment using simulation studies by considering several different models with a range of unknown parameters and hyperparameters. Our simulations from the previous chapters show similar performances of the methods under uniform and sparse priors, we only evaluate our methods assuming a uniform prior for the balance indicate vector $\mathbf{z}$. We expect similar conclusions using the sparse prior.

4.4.1. Data generation

The simulated data are generated according to model illustrated in Figure 4.1, where n sample are divided equally into two treatment groups. With the prespecified values for $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma}$, we generate $\boldsymbol{p}_i$ for each subject. The observed count $\boldsymbol{x}_i$ is simulated from a multinomial distribution with the total counts being equal for all the subjects. The outcome is generated by calculating the balance index with $\boldsymbol{p}_i$ and $\boldsymbol{z}$.

We consider various combinations for true parameter values. We generate data for a total sample size $n = 100$ with the compositional vector of dimension $q = 10$. The balance indicator vector $\boldsymbol{z}$ has value 1 for the first 2 coordinates and -1 for the next two coordinate; the rest are all 0. For the first three models (Model 1 - Model 3), we assume that the first 4 coordinates in mean parameters of the logistic normal distribution in the untreated and treated group are $\boldsymbol{\mu_1} = (1.4, 1.3, 1.2, 1.10$ and $\boldsymbol{\mu_0} = (1.1, 1.2, 1.3, 1.4)$, respectively, with the values in other components of the two mean parameters being randomly generated from the standard normal distribution. We consider such a setting in simulations in order to test how the unrelated values in compositional vector affect the estimation. This setting simulates the scenario that the treatment affects microbial composition, including the components that are nor relevant to the balance. The variance-covariance matrix $\boldsymbol{\Sigma}$ is set to identity matrix with dimensionality $q - 1$. The path coefficients b_z, c_zin the second equation are set to be 1.0, respectively. Finally, the random error is assumed to be normally distributed with mean 0 and variance 1.

For the next three models, Model 4 - Model 6, we assume $\boldsymbol{\mu}_1 = \boldsymbol{\mu_0} = (1.1, 1.2, 1.3, 1.4)$ and all other parameters are the same as in the previous models. This setting simulates the scenario that the treatment does not affect the microbial composition, however, microbiome composition is

associated with the continuous outcome through balance. Table 4.1 shows the effects of balance on the outcome and the mediation effect for each of these 6 models.

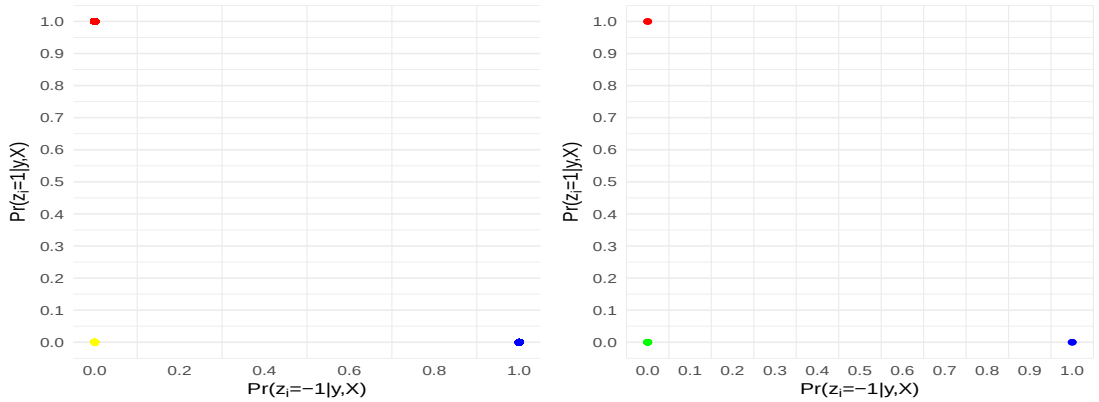
Table 4.1: Parameters used in the simulations. True values for the first four components in the population mean vector for the compositions in treated and untreated group, effect of balance on outcome and mediation effect.

	μ_0	μ_1	bal($\boldsymbol{a}, \boldsymbol{z}$)	b_z	mediation effect
Model 1	$(1.1, 1.2, 1.3, 1.4)$	$(1.4, 1.3, 1.2, 1.1)$	0.4	1	0.4
Model 2	$(1.1, 1.2, 1.3, 1.4)$	$(1.4, 1.3, 1.2, 1.1)$	0.4	0.5	0.2
Model 3	$(1.1, 1.2, 1.3, 1.4)$	$(1.4, 1.3, 1.2, 1.1)$	0.4	0	0
Model 4	$(1.1, 1.2, 1.3, 1.4)$	$(1.1, 1.2, 1.3, 1.4)$	0	1	0
Model 5	$(1.1, 1.2, 1.3, 1.4)$	$(1.1, 1.2, 1.3, 1.4)$	0	0.5	0
Model 6	$(1.1, 1.2, 1.3, 1.4)$	$(1.1, 1.2, 1.3, 1.4)$	0	0	0

4.4.2. Simulation results for models with mediation effect

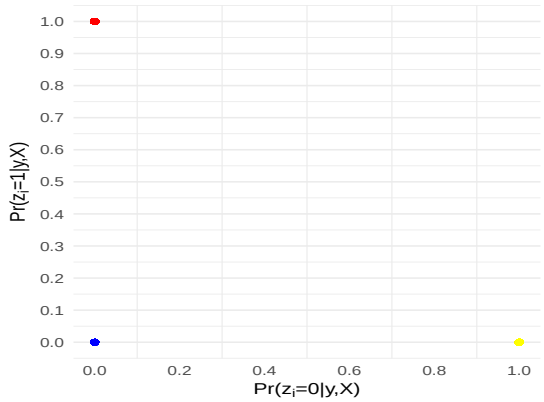
Figure 4.2, Figure 4.3 and Figure 4.4 shows the results across 100 replications for the models 1-3, where the true μ_0 and μ_1 differ in the first four components and the treatment has effects on microbiome composition.

For models with large or intermediate mediation effects, the MCMC algprithm is able to identify the balance configurations $\boldsymbol{z}$ well, with the marginal posterior probability being the highest in the z_+ for the first 2 components, and highest in the z_- set for the 3rd and 4th components. The posterior mean of mediation effect is also centered around the true values. With $b_z = 0.5$, we can correctly recover the first four components of $\boldsymbol{z}$ but we also see that some taxa in the z_0 set can have highest posterior probability of being in the z_+ group. With $b_z = 0$, it is impossible to correctly recover $\boldsymbol{z}$. This observation is consistent with the updating step in MCMC, where only outcome $\boldsymbol{Y}$ and latent relative matrix $\boldsymbol{P}$ are used to find the next sample of $\boldsymbol{z}$. As $\boldsymbol{Y}$ contain less information with decreased b_z values, the algorithm is not able to find balance configuration vector $\boldsymbol{z}$.

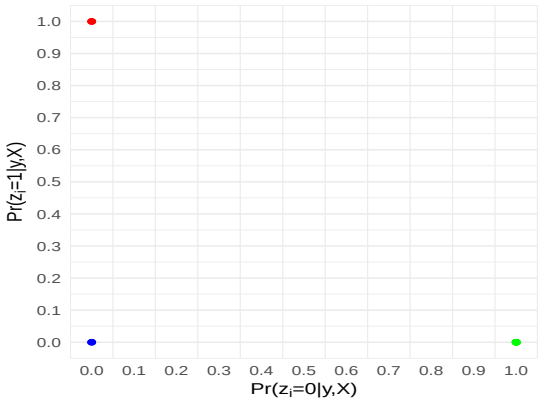


(a) all simulation points

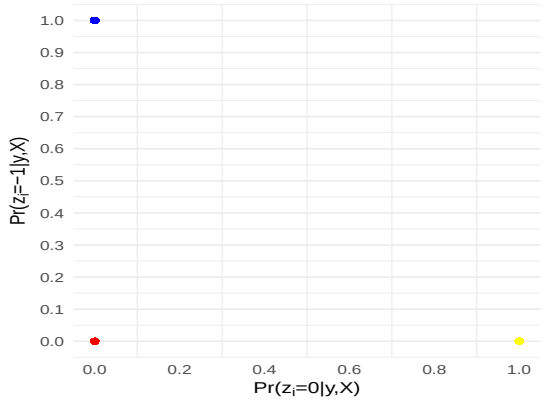
(b) averaged across simulations



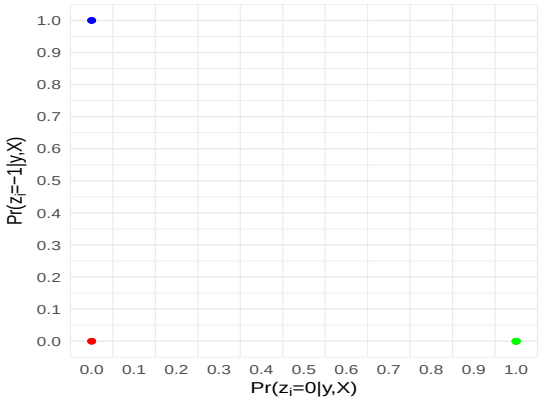
(c) all simulation points



(d) averaged across simulations

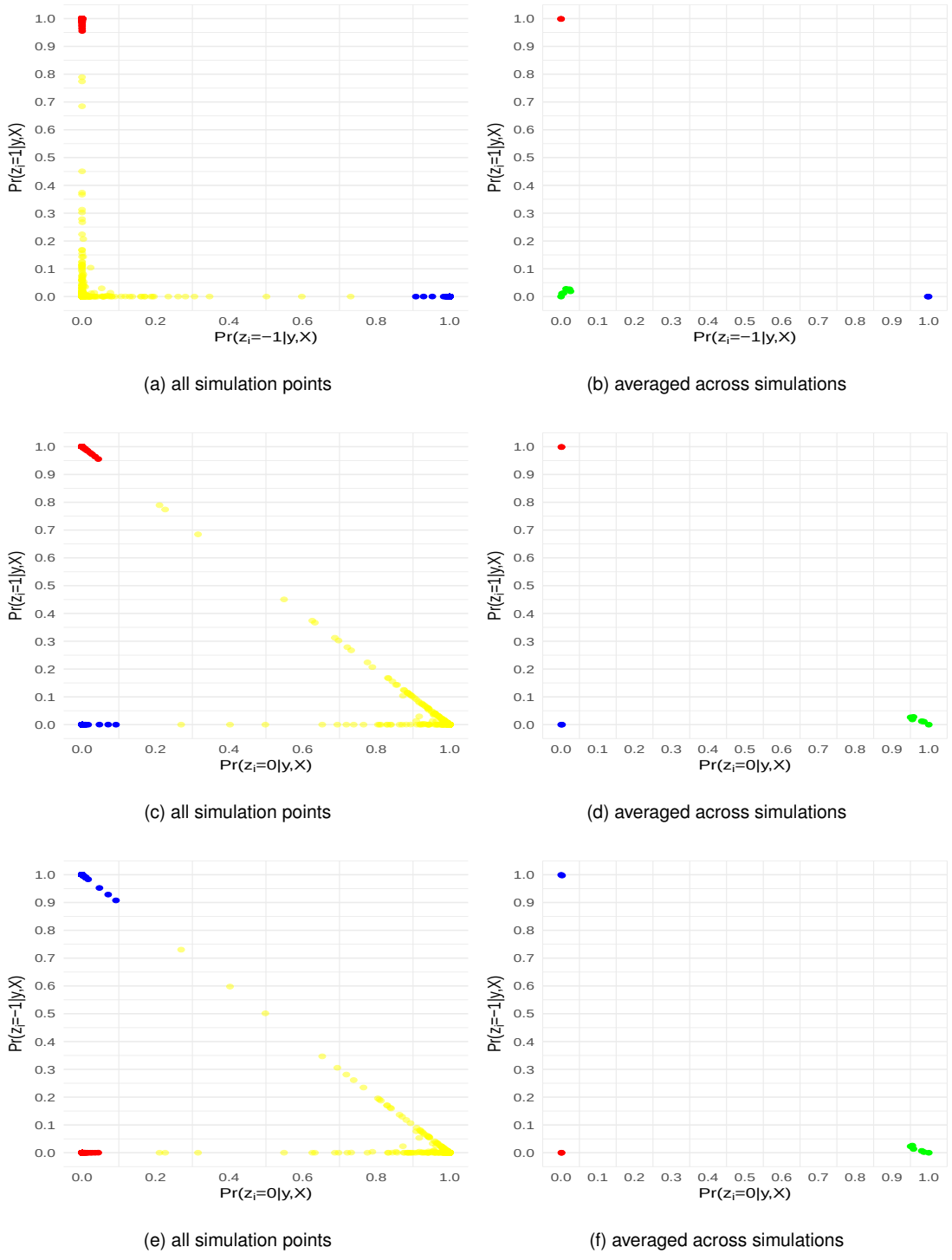


(e) all simulation points



(f) averaged across simulations

Figure 4.2: Simulation results for balance mediation analysis in Model 1 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).



(a) all simulation points

(b) averaged across simulations

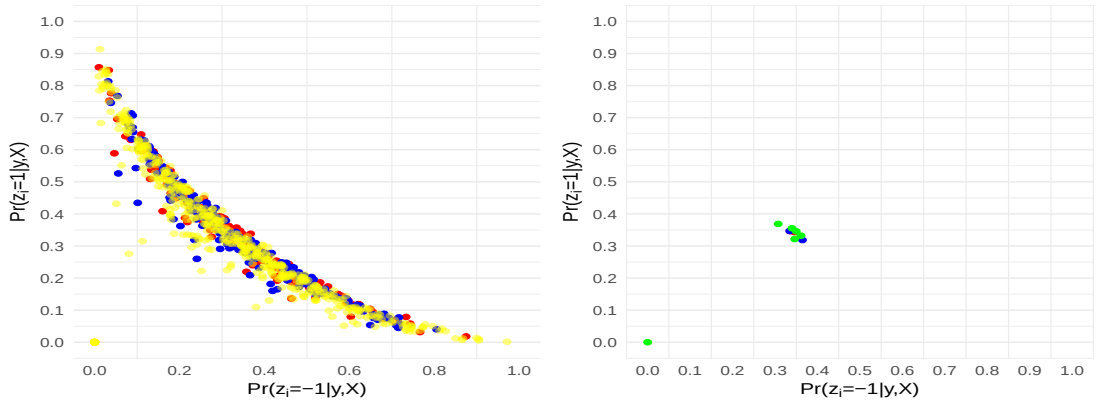
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

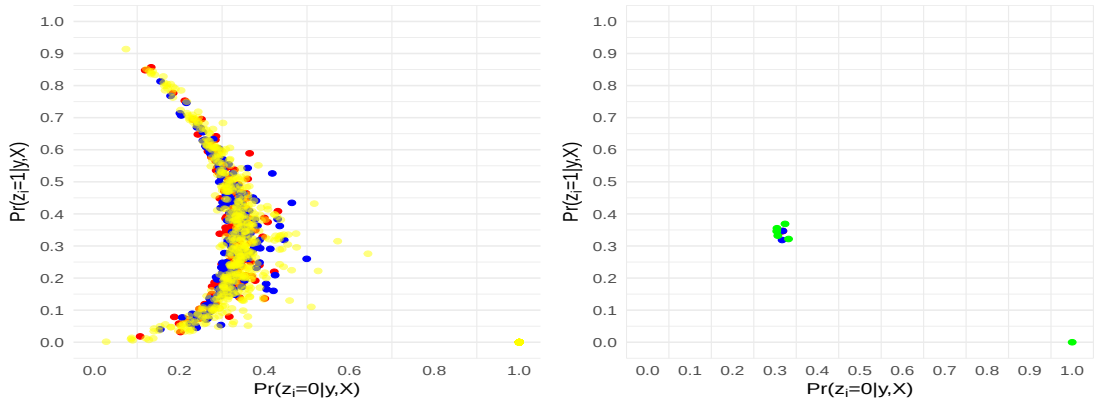
(f) averaged across simulations

Figure 4.3: Simulation results for balance mediation analysis in Model 2 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).



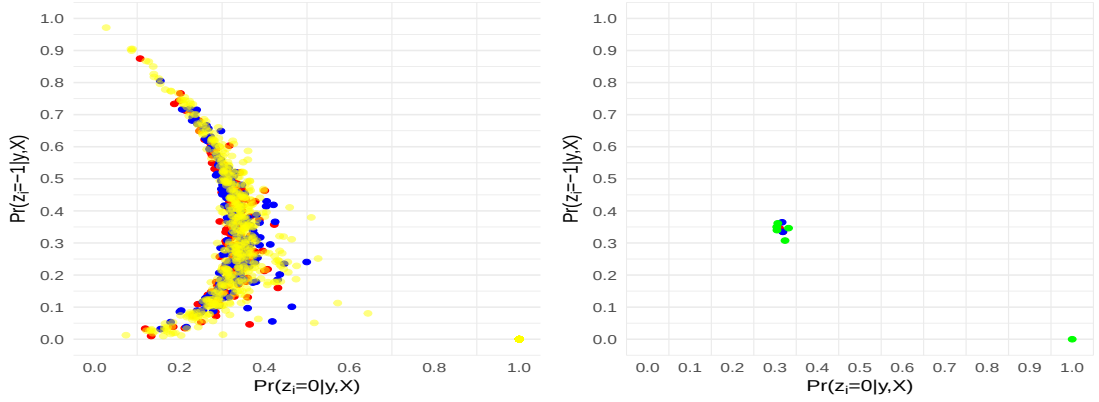
(a) all simulation points

(b) averaged across simulations



(c) all simulation points

(d) averaged across simulations



(e) all simulation points

(f) averaged across simulations

Figure 4.4: Simulation results for balance mediation analysis in Model 3 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).

4.4.3. Simulation results for models with no mediation effect

Figure 4.5, Figure 4.6 and Figure 4.7 shows similar results for models 4-6, where μ_0 and μ_1 are the same and the treatment does not affect the microbiome composition. However, note that only the outcome regression provides information for recovering the balance configurations.

(a) all simulation points

(b) averaged across simulations

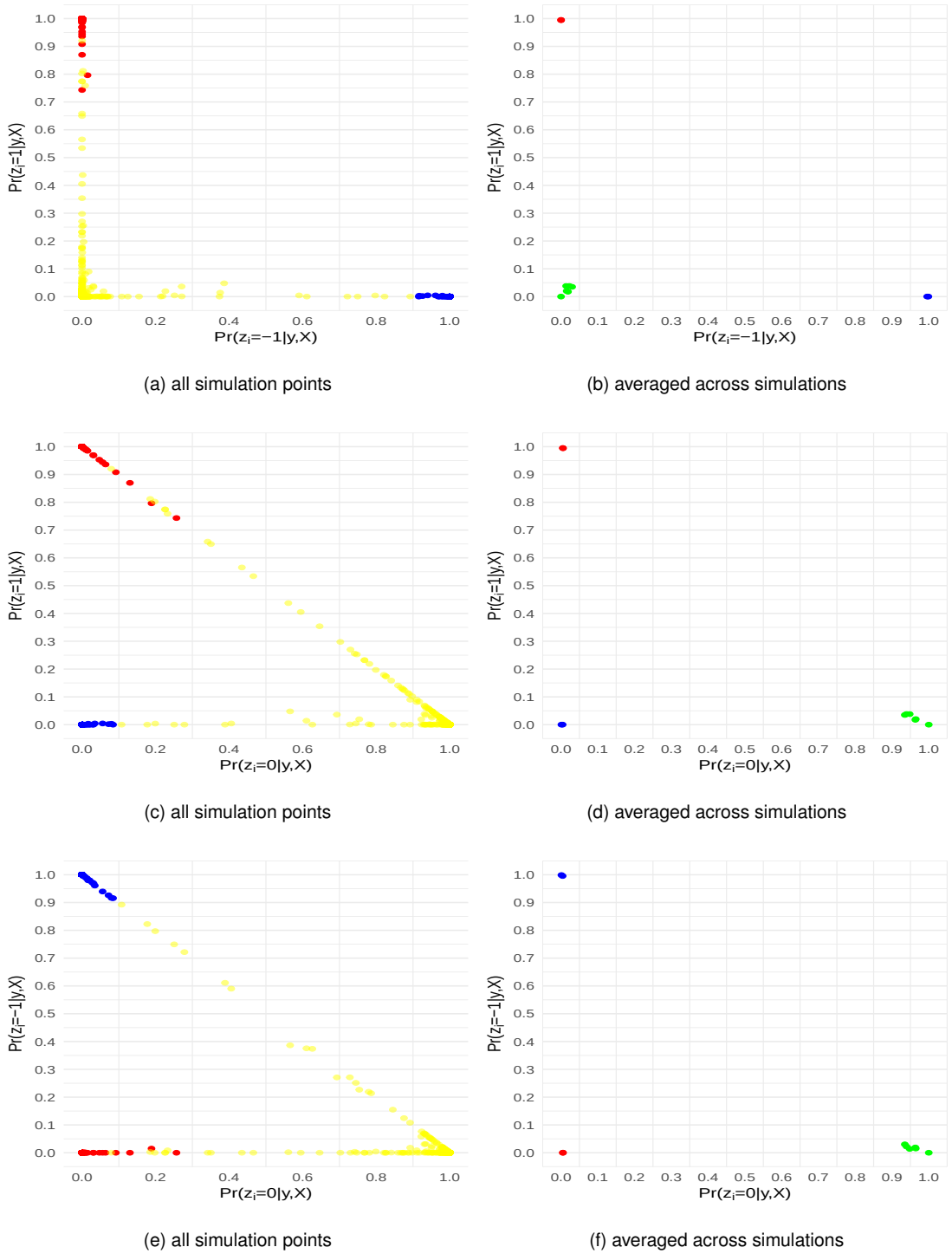
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

(f) averaged across simulations

Figure 4.5: Simulation results for balance mediation analysis in Model 4 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).



(a) all simulation points

(b) averaged across simulations

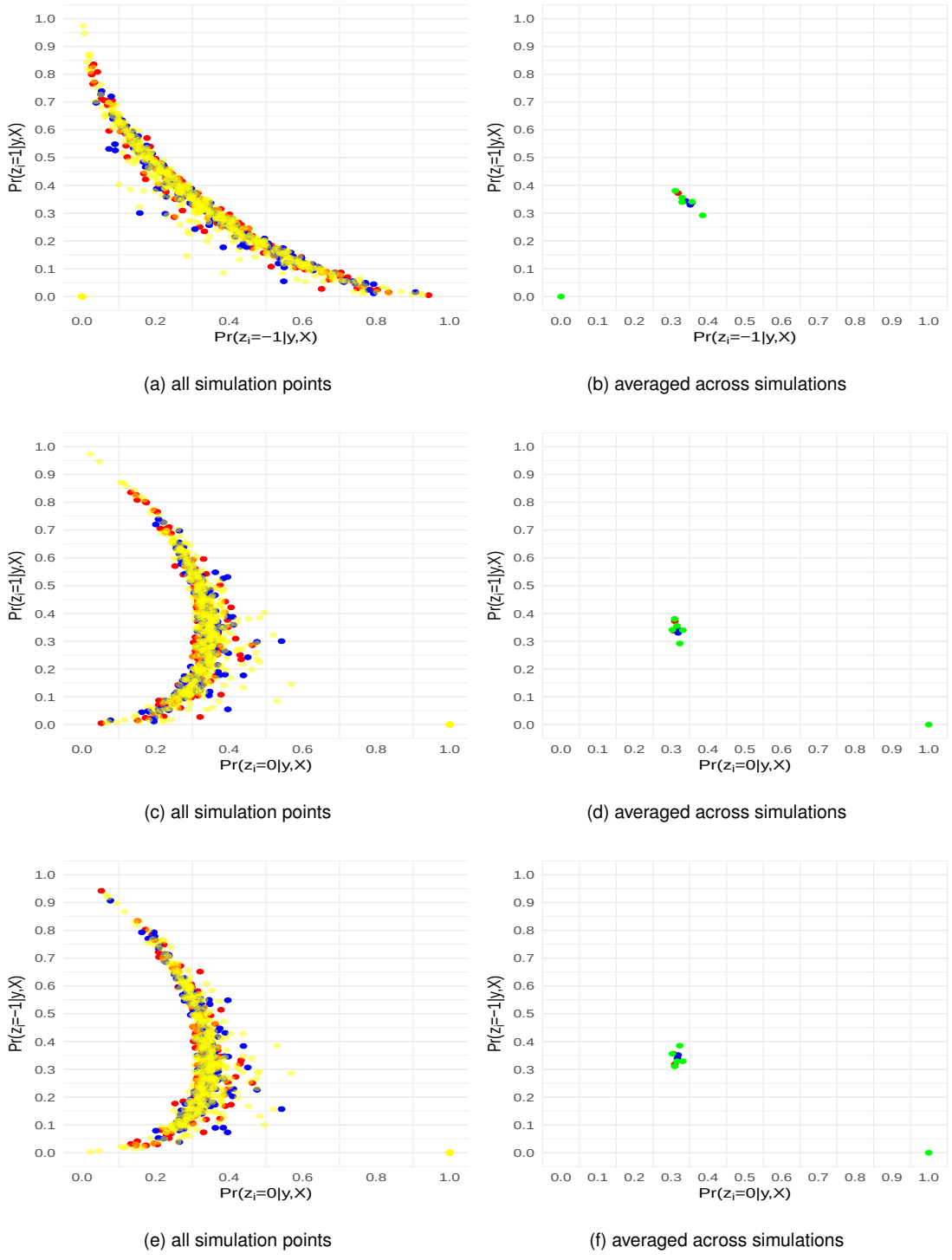
(c) all simulation points

(d) averaged across simulations

(e) all simulation points

(f) averaged across simulations

Figure 4.6: Simulation results for balance mediation analysis in Model 5 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).



(a) all simulation points

(b) averaged across simulations

(c) all simulation points

(d) averaged across simulations

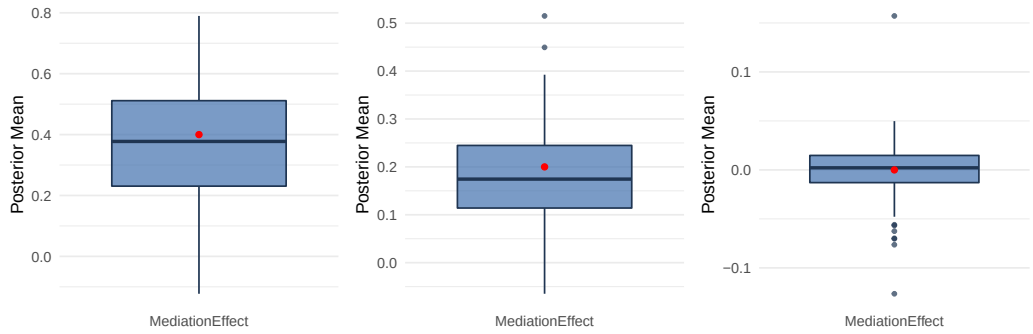
(e) all simulation points

(f) averaged across simulations

Figure 4.7: Simulation results for balance mediation analysis in Model 6 under uniform prior for z. Pairwise posterior probabilities with all 100 simulations and averaged over 100 simulations are plotted for taxa in the positive set z_+ (red), negative set z_- (blue) and null set z_0 (yellow/green).

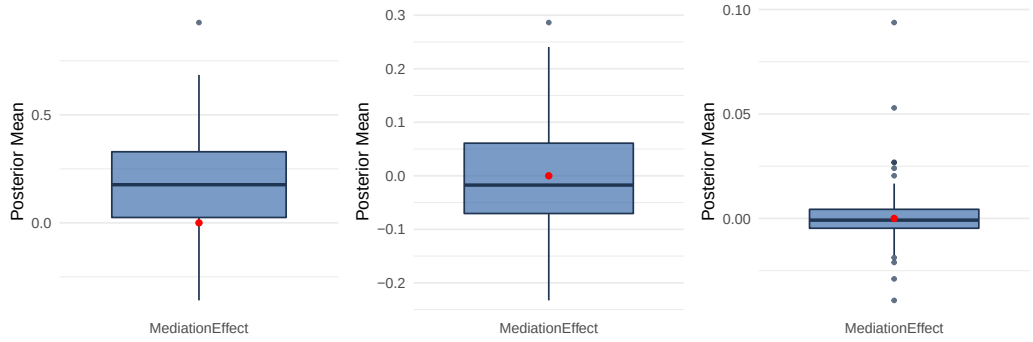
4.4.4. Posterior inference of mediation effect

The estimation of posterior mean of the mediation effect (Figure 4.8 and Figure 4.9)has a positive bias when $b_z = 0$ and the corresponding mediation effect is zero. This is because during MCMC sampling for the purpose of identifiability of the balance configuration, we choose the next sampling value of the balance configuration vector in such a way that the resulting regression coefficient b_z is always positive.



(a) $\beta_{1z} = 1$ (b) $\beta_{1z} = 0.5$ (c) $\beta_{1z} = 0$

Figure 4.8: Boxplots of the posterior mean of the mediation effect over 100 simulations for Model 1 to Model 3.



(a) $\beta_{1z} = 1$ (b) $\beta_{1z} = 0.5$ (c) $\beta_{1z} = 0$

Figure 4.9: Boxplots of the posterior mean of the mediation effect over 100 simulations for Model 4 to Model 6.

4.5. Applications to mediation analysis of plasma metabolomics data

4.5.1. Summary of serum metabolites

Wu et al., 2016 reported a study that used 16S rRNA-tagged sequencing as well as plasma and urinary metabolomic platforms, to compare the measures of dietary intake, gut microbiota composition and the plasma metabolome between 15 healthy human vegans and 16 omnivores, sampled in an urban USA environment. One important question is to understand how diet affects plasma metabolites and how much such effects are mediated through gut microbiome.

Compared to urine metabolites, we focus on analyzing serum metabolites since it is more related to diet and gut microbiome. The data set includes 349 serum metabolites that belong to 9 pathways: amino acid, carbohydrate, cofactors and vitamins, energy, lipid, nucleotide, peptide, secondary metabolism, and xenobiotics. Due to the limited sample sizes, we select the following serum metabolites in our analyses: 4-ethylphenyl sulfate, 3-hydroxyhippurate,hippurate and catechol sulfate. These four metabolites belong to the same Xenobiotics Benzoate metabolism pathway and are statistically significantly different among two diet groups (Figure 4.10). They are also the products of gut microbiota, so it is of interest to dissect how much of the difference in these metabolites between the two diet groups attributes to microbiota. By restricting our analysis to only a few metabolites, we also decrease the chance of false discoveries that are associated with a large number of tests.

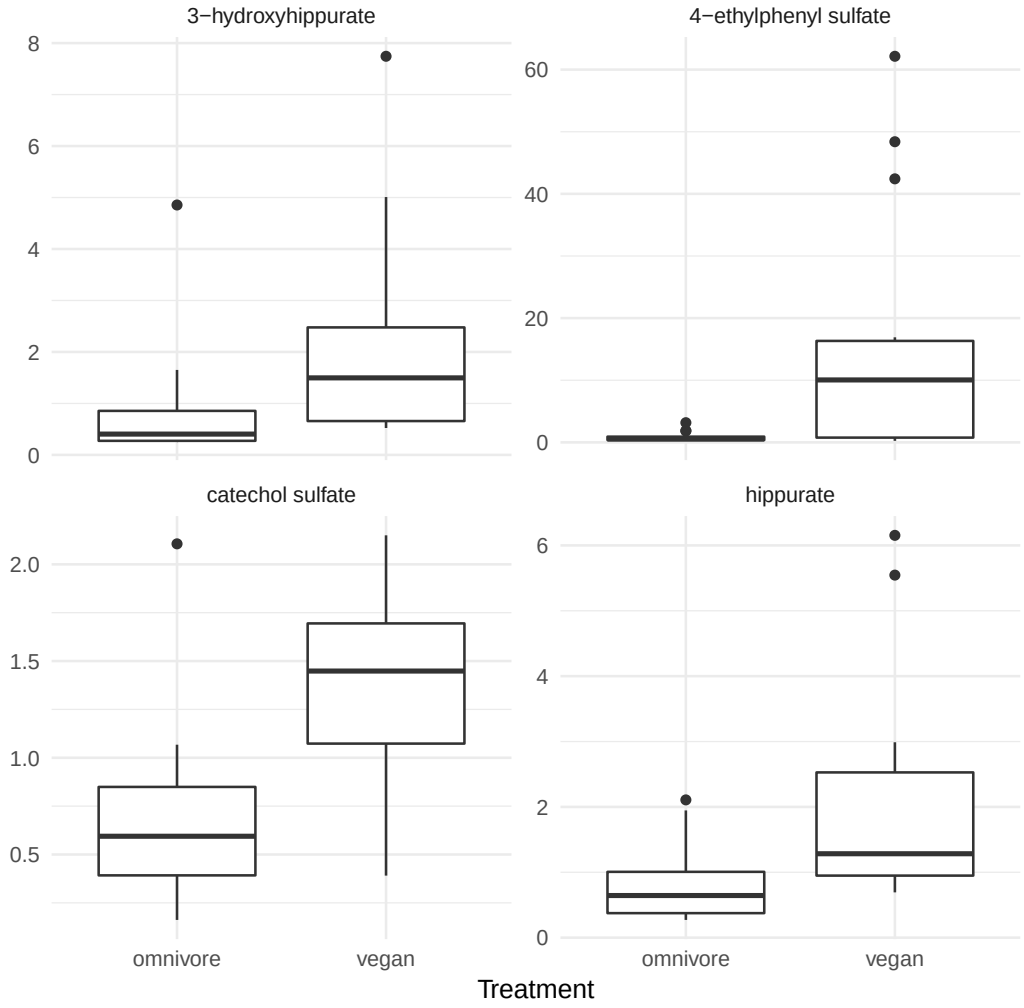
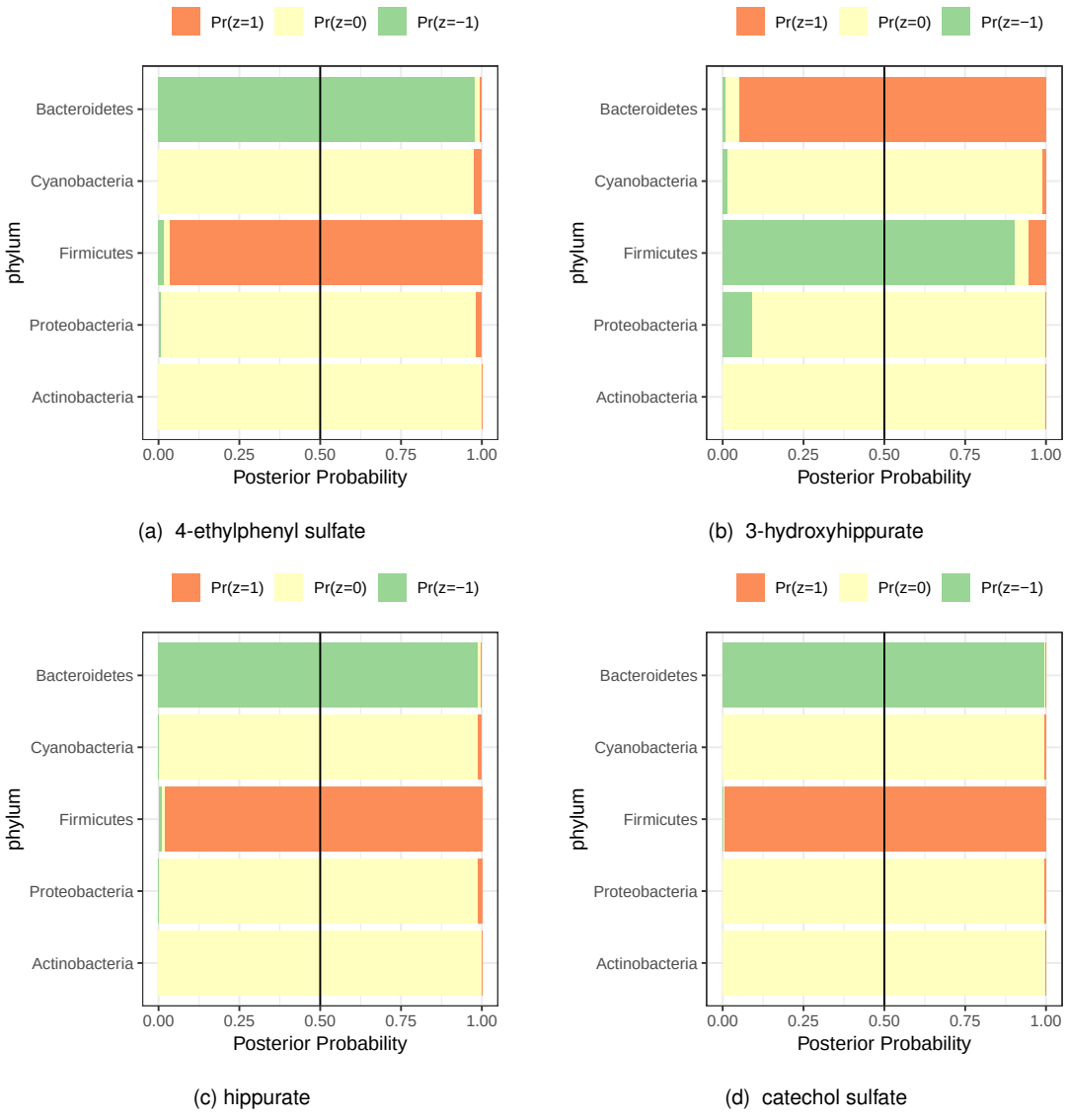


Figure 4.10: Boxplots of four selected metabolites for each diet group. The between group differences are statistically significant with $p < 0.05$.

4.5.2. Mediation analysis at the phylum level

Our first analysis is used as a proof-of-concept of our proposed methods, where we summarize the bacterial count data at the phylum level. For each sample, we have read counts for each of the five different phyla, including *Actinobacteria, Bacteroidetes,Cyanobacteria, Firmicutes,* and *Proteobactreria*. In our analysis, we choose *Proteobactreria* as the reference phylum in the logistic-normal distribution in modeling the taxon composition. Figure 4.11 shows the plot of the posterior distribution of the balance configuration of four bacterial phyla in Bayesian balance mediation analysis of the effect of vegan diet on the selected serum metabolites under sparse prior The posterior plots

indicate that the ratio of phyla *Firmicutes* and *Bacteroides* defines the balance that mediates the effect of vegan diet on all six metabolites. Indeed, the ratio of the relative abundance of phylum *Firmicutes* and phylum *Bacteroides* has been shown to be associated with fat intake and also diet in multiple previous publications (Holmes et al., 2011; Tomova et al., 2019). Figure 4.12 shows the posterior mean of the relative abundance of each phylum in the two diet groups, indicating a weak effect of diet on phylum composition.



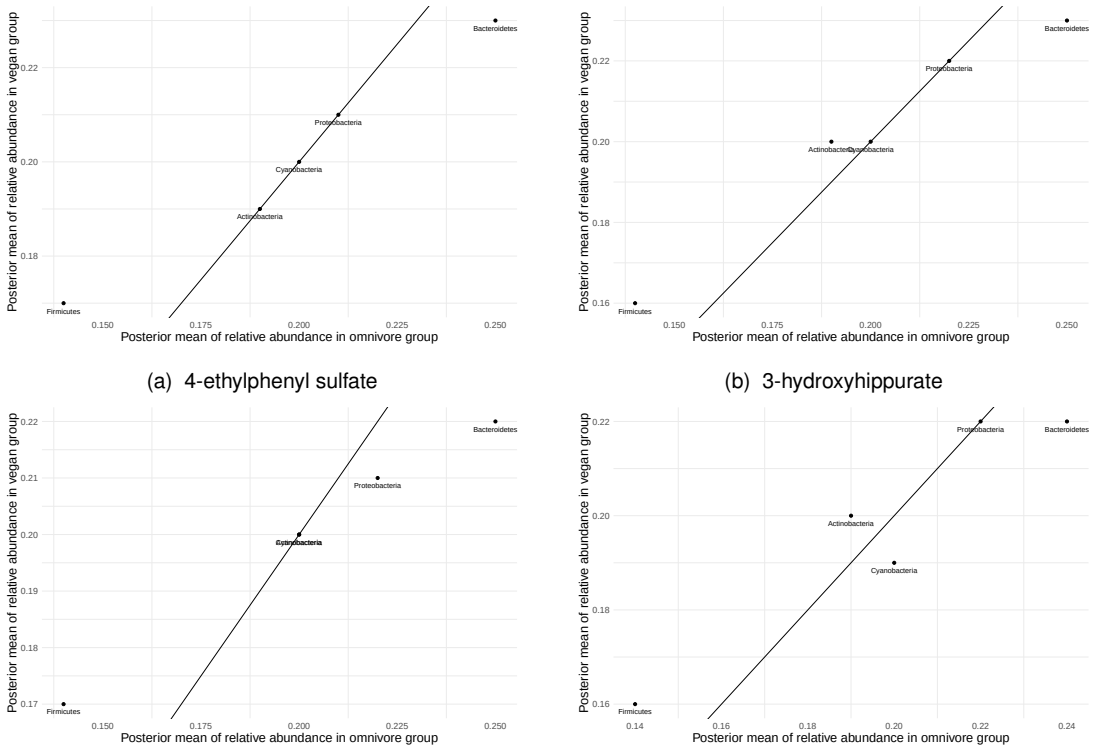
(a) 4-ethylphenyl sulfate

(b) 3-hydroxyhippurate

(c) hippurate

(d) catechol sulfate

Figure 4.11: Plot of the posterior distribution of the balance configuration of five bacterial phyla in Bayesian balance mediation analysis of the effect of vegan diet on different serum metabolites under sparse prior for z.



(a) 4-ethylphenyl sulfate

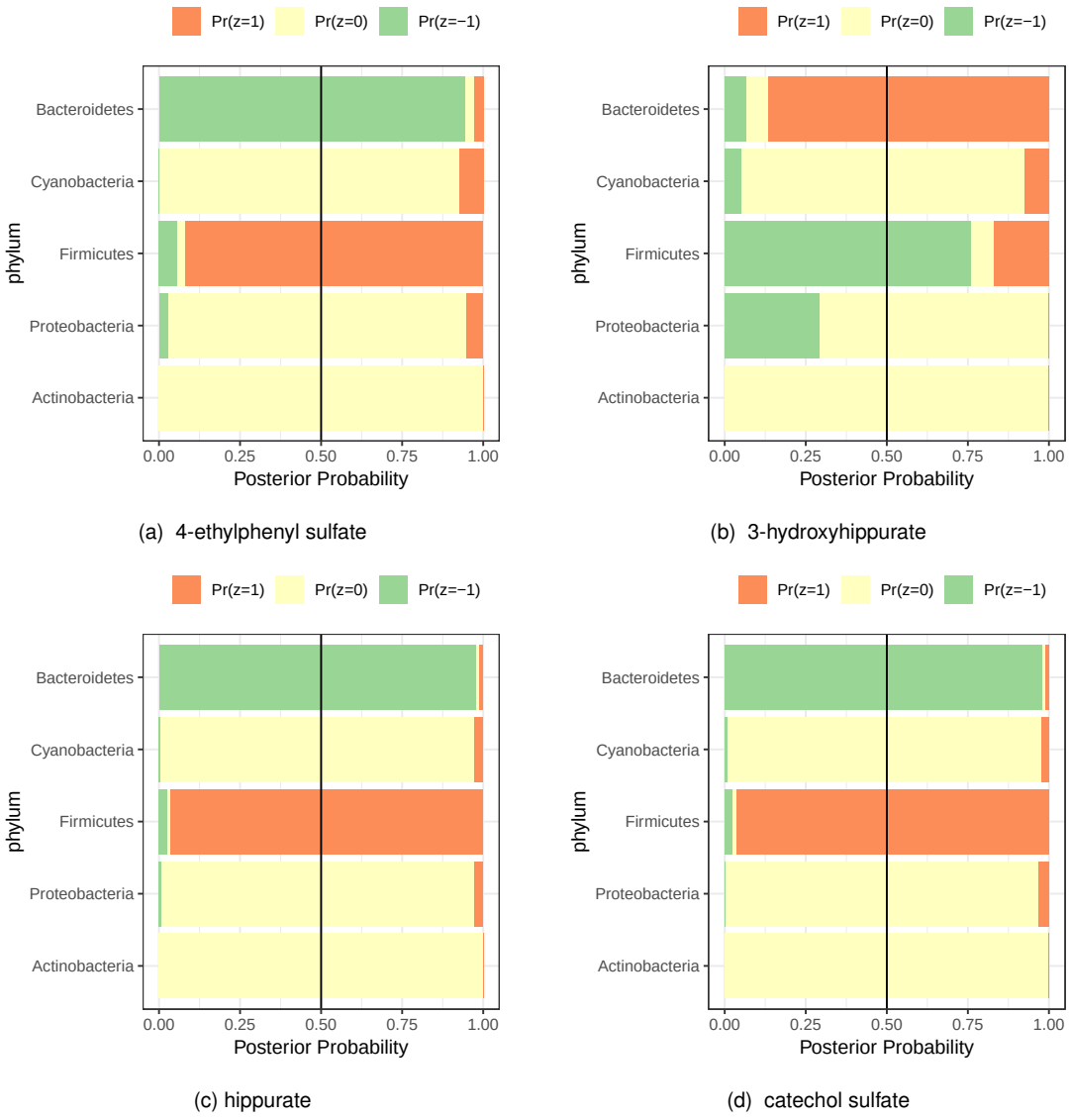
(b) 3-hydroxyhippurate

(c) hippurate

(d) catechol sulfate

Figure 4.12: Plot of the posterior mean of the relative abundance of each phylum under the sparse prior of z. Solid line represents equal proportions in vegan and omnivore diet group.

Under the uniform prior, the configuration of identified balance are similar to those under sparse prior (Figure 4.13) and the estimated relative abundance are also similar (Figure 4.14). In some scenarios, more phyla are included in the balance. This is expected because with uniform prior the assumption is the almost 1/3 of the phyla contribute to balance.



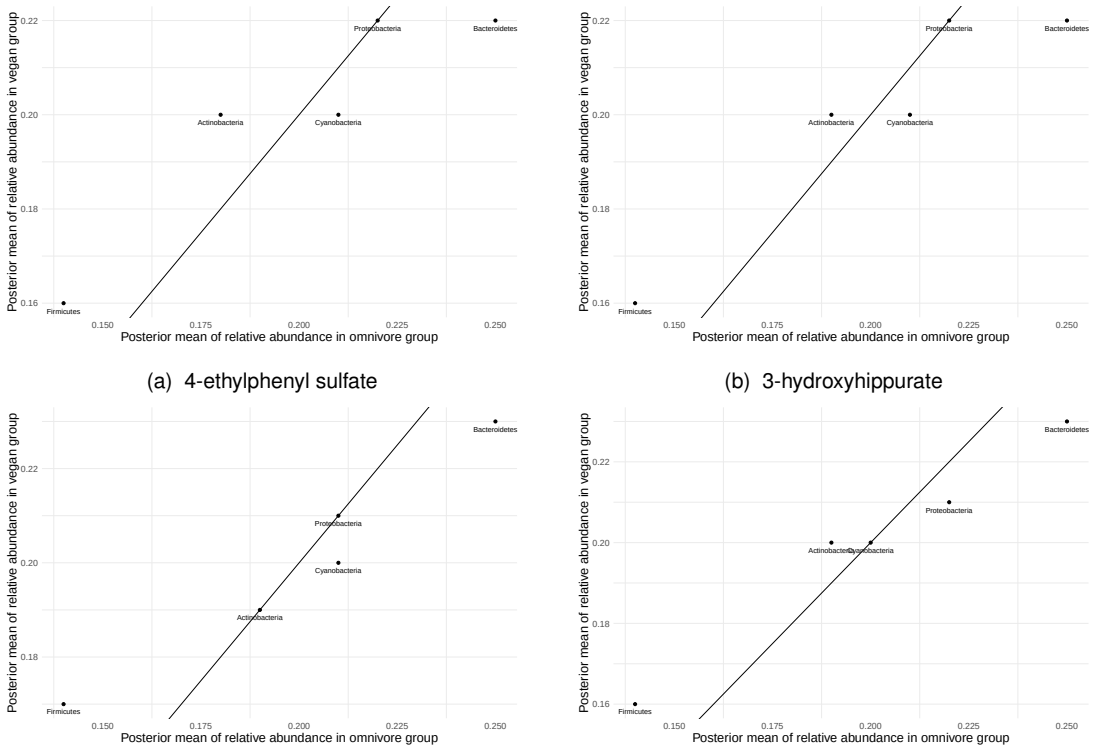
(a) 4-ethylphenyl sulfate

(b) 3-hydroxyhippurate

(c) hippurate

(d) catechol sulfate

Figure 4.13: Plot of the posterior distribution of the balance configuration of five bacterial phyla in Bayesian balance mediation analysis of the effect of vegan diet on different serum metabolites under uniform prior for z.



(a) 4-ethylphenyl sulfate

(b) 3-hydroxyhippurate

(c) hippurate

(d) catechol sulfate

Figure 4.14: Plot of the posterior mean of the relative abundance of each phylum under the uniform prior of z. Solid line represents equal proportions in vegan and omnivore diet group.

Table 4.2 shows that estimated direct effect of vegan diet on each metabolite and the mediation effect of the identified balance and their 95 credible intervals. We observe a strong direct effect of vegan diet on each of the metabolites and also moderate effect of the identified balance on metabolites, but a relatively weak mediating effect of the balance. This is largely explained by the relatively weak effect of diet on microbiome composition (see Figures 4.12 and 4.14).

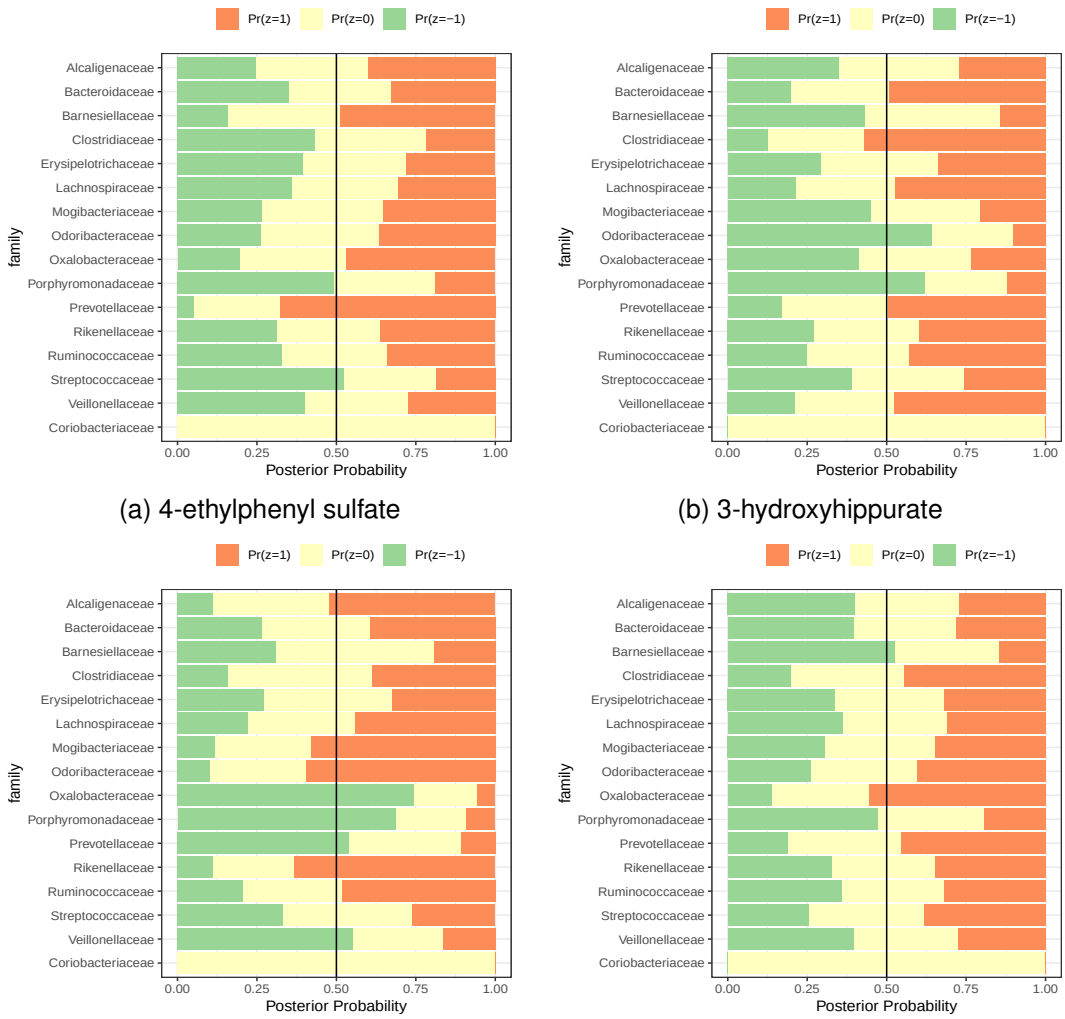
One interesting observation in our phylum level analysis is that balance configuration is reversed for 3-hydroxyhippurate compared to the other three metabolites. This indicates that there is a negative correlation between 3-hydroxyhippurate and the other three. We further examined the data and observed that such a negative correlation is more prominent in subjects with vegan diet, indicating that the balance identified in our analysis is indeed mediating the effect of vegan diet on different metabolites. However, due to the limited sample size of our data, the effect did not reach statistical significance.

Table 4.2: Results of Bayesian balance mediation analysis at the phylum level under uniform and sparse priors. The table shows the estimation of direct, mediation effects on vegan diet on different metabolites and the effect of balance on each metabolite, based on the predictive posterior distributions.

	Uniform prior			
	4-ethylphenyl sulfate	3-hydroxyhippurate	hippurate	catechol sulfate
Vegan (T)	14.83(14.2,15.73)	1.1(0.89,1.45)	1.33(1.23,1.48)	0.73(0.71,0.77)
Balance	1.18(0,3.53)	0.05(0,0.37)	0.42(0,0.91)	0.24(0,0.38)
Mediation	0.22(-0.47, 1.58),	-0.01 (-0.12 , 0.02)	0.09 (-0.13, 0.50)	0.04 (-0.09,0.21)
	Sparse prior			
	4-ethylphenyl sulfate	3-hydroxyhippurate	hippurate	catechol sulfate
Vegan(T)	14.67(14.48,16.01)	1.18(0.97,1.25)	1.28(1.26,1.37)	0.72(0.71,0.76)
Balance	2.54(0.01,3.69)	0.23(0,0.42)	0.8(0,0.92)	0.33(0,0.38)
Mediation	0.51(-0.71,1.95)	-0.04(-0.21,0.07)	0.18(-0.18,0.56)	0.06(-0.09,0.23)

4.5.3. Analysis at the family level

Our second analysis is performed at the family level, where we summarize the bacterial counts data at the family level, and for each sample, we have read counts for each of the 16 different families, including *Alcaligenaceae, Bacteroidaceae, Barnesiellaceae, Clostridiaceae, Coriobacteriaceae, Erysipelotrichaceae, Lachnospiraceae, Mogibacteriaceae, Odoribacteraceae, Oxalobacteraceae, Porphyromonadaceae, Prevotellaceae, Rikenellaceae, Ruminococcaceae, Streptococcaceae* , and *Veillonellaceae*. In our analysis, we choose *Clostridiales* as the reference family in the logistic-normal distribution. Figure 4.15 shows the plot of the posterior distribution of the balance configuration of bacterial families in Bayesian balance mediation analysis of the effect of vegan diet on selected serum metabolites, where uniform prior is assumed. With family level data, the balance defined in the mediation analysis differ among different metabolites. More specifically, ratio between *Barnesiellaceae, oxalobacteraceae* and *clostridiaceae* defines the balance that mediates the effect of vegan diet on 4-ethylphenyl sulfate. For 3-hydroxyhippurate, the balance is defined by the ratio between *Bacteroidaceae,Clostridiaceae,Lapchnospiraceae,Mogibacteriaceae,Veillonellaceae* and *Erysipelotrichaceae, Odoribacteraceae, Porphyromonadaceae*. For hippurate, balance is defined by *Lachnospiraceae, Mogibacteriaceae, Oxalobacteraceae* and *Porphyromonadaceae, Prevotel-*



(a) 4-ethylphenyl sulfate

(b) 3-hydroxyhippurate

(c) hippurate

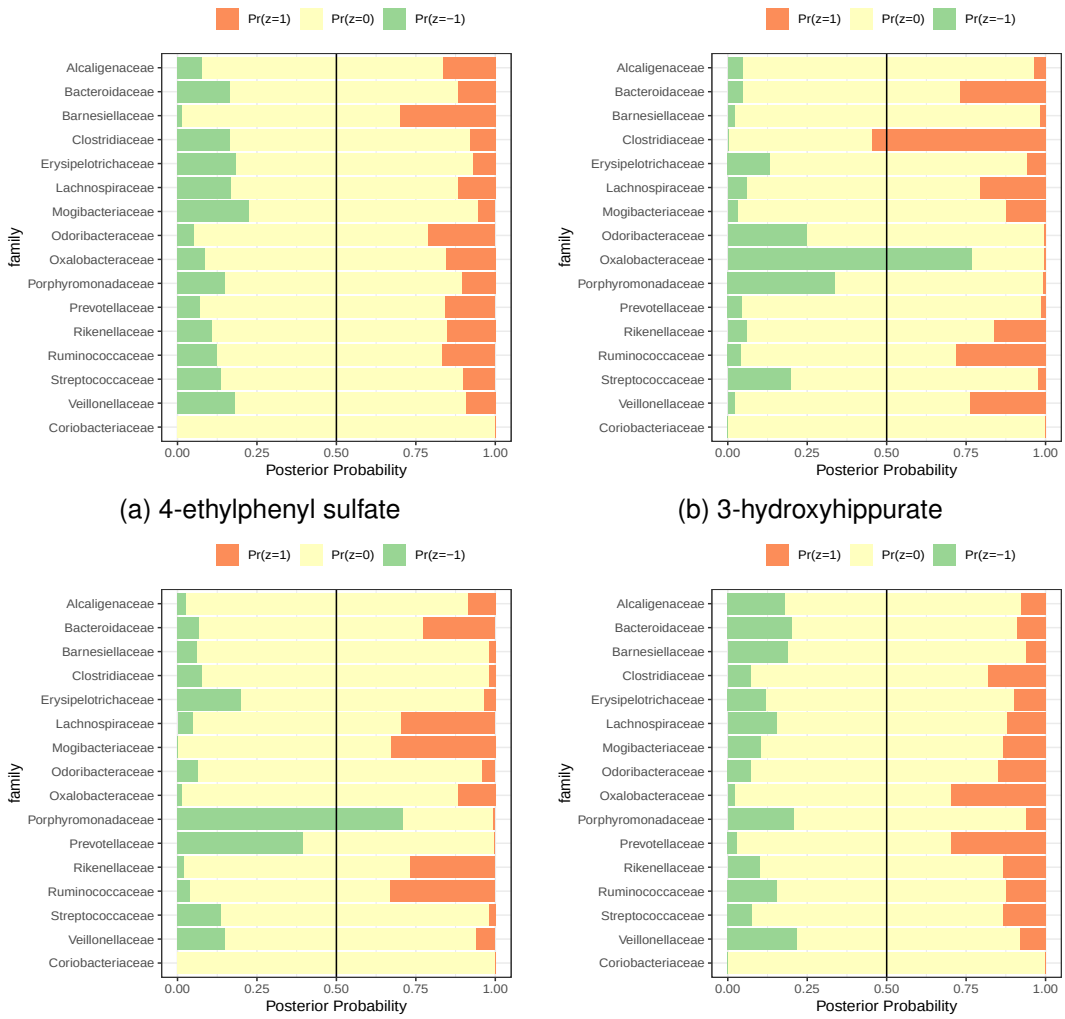
(d) catechol sulfate

Figure 4.15: Plot of the posterior distribution of the balance configuration of 16 bacterial families in Bayesian balance mediation analysis of the effect of vegan diet on different serum metabolites under a uniform prior for z.

laceae, Veillonellaceae. For catechol sulfate, no clear balance can be identified that mediate the effect of vegan diet. The balance identified at the family level differs among different metabolites indicates that different groups of bacteria involve in the generation of these metabolites, which is supported by multiple publications(Besten et al., 2013; Morrison and Preston, 2016).

Under the sparse prior, no clear balance can be identified for these metabolites (see Figure 4.16) except for 3-hydroxyhippurate. This is largely due to the small sample sizes in our study.

Table 4.3 shows that estimate direct effect of vegan diet on different metabolites and the mediation



(a) 4-ethylphenyl sulfate

(b) 3-hydroxyhippurate

(c) hippurate

(d) catechol sulfate

Figure 4.16: Plot of the posterior distribution of the balance configuration of 16 bacterial families in Bayesian balance mediation analysis of the effect of vegan diet on different serum metabolites under a sparse prior for z.

effect by the identified balance and their 95 credible intervals under both uniform and sparse priors. The estimated direct effect and mediation effect are similar to our analysis at the phylum level but the credible intervals are wider, indicating greater uncertainty. The identified balance is associated with the metabolite levels, but there is no mediation effect of the identified balance.

Table 4.3: Results of Bayesian balance mediation analysis of four serum metabolites at the family level under uniform for z and sparse priors. The table shows the estimation of direct, mediation effects on vegan diet on different metabolites and the effect of balance on each metabolite, based on the predictive posterior distributions.

		Uniform prior		
	4-ethylphenyl sulfate	3-hydroxyhippurate	hippurate	catechol sulfate
Vegan(T)	14.48(12.35,16.71)	1.16(0.66,1.57)	1.43(0.95,1.93)	0.70(0.61,0.78)
Balance	1.09(0.11,2.65)	0.26(0.04,0.5)	0.23(0.03,0.46)	0.04(0,0.09)
Mediation	0.03(-1.71,1.87)	0.02(-0.39,0.45)	-0.04 (-0.46,0.36)	0(-0.07,0.07)
		Sparse prior		
	4-ethylphenyl sulfate	3-hydroxyhippurate	hippurate	catechol sulfate
Vegan(T)	13.95(11.41,16.07)	1.19(0.84,1.62)	1.49(1.24,1.73)	0.71(0.62,0.79)
Balance	0.90(0.07,2.44)	0.24(0.08,0.49)	0.22(0.04,0.46)	0.04(0,0.11)
Mediation	0.17 (-1.33,2.14)	0.03 (-0.32, 0.38)	-0.07 (-0.40, 0.21)	0(-0.06,0.07)

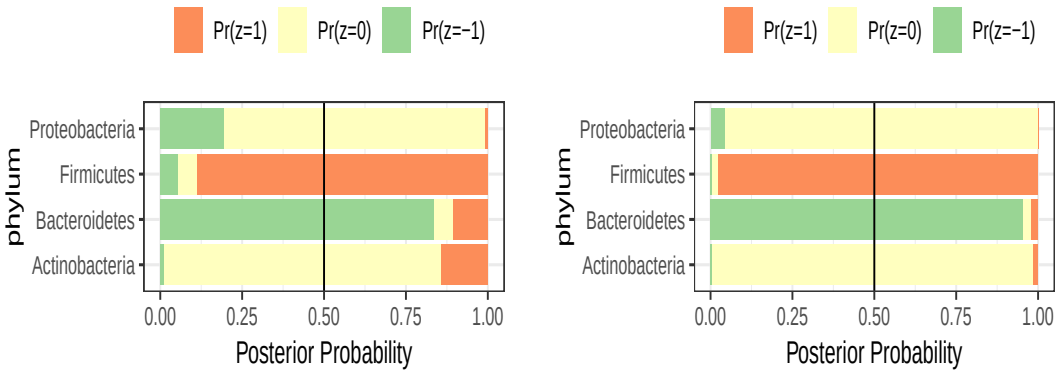
4.6. Application to mediation analysis of COMBO data

We also apply the Bayesian balance mediation analysis to a cross-sectional study of 98 healthy volunteers as reported in Wu et al., 2011. The dataset consists of 16SrRNA sequences from fecal samples of 98 healthy individuals from the University of Pennsylvania. It also contains demographic and clinical information including fat intake and BMI, where the habitual long-term fat intake was derived from the food frequency questionnaire (FFQ). Such measurements are widely applied in nutritional research, and their reproducibility and validity have been validated (Hu et al., 1999). We summarized operational taxonomic units (OTUs) at the genus level and then filtered out the genera that appear in fewer than 10% of the samples, leaving 45 genera in 98 samples. These 45 genera can be further grouped into 5 phyla and 8 orders.

We apply the proposed Bayesian mediation analysis to investigate the mediating effect of gut microbiome in linking high fat intake and BMI.

4.6.1. Mediation analysis of the COMBO data at the phylum level

Our first analysis is used as a proof-of-concept of our proposed methods, where we summarize the bacterial count data at the phylum level. For each sample, we have read counts for each of the five different phyla, including *Actinobacteria, Bacteroidetes, Firmicutes,* and *Proteobactreria*. In our analysis, we choose *Proteobactreria* as the reference phylum in the logistic-normal distribution in modeling the taxon composition. Figure 4.17 shows the plot of the posterior distribution of the balance configuration of four bacterial phyla in Bayesian balance mediation analysis of the effect of high-fat on BMI under both uniform prior and sparse priors. Since the number of phyla is small, we expect that both uniform-prior and sparse-prior give similar results. The posterior plots indicate that the ratio of phyla *Firmicutes* and *Bacteroides* defines the balance that mediates the effect of high fat on BMI. Indeed, the ratio of the relative abundance of phylum *Firmicutes* and phylum *Bacteroides* has been shown to be associated with fat intake and also the obesity in multiple previous publications (Ley et al., 2006; Wu et al., 2011).



(a) uniform prior

(b) sparse prior

Figure 4.17: The posterior distributions of the balance configuration of four bacterial phyla in Bayesian balance mediation analysis of the effect of high-fat on BMI under uniform and sparse prior for z, respectively.

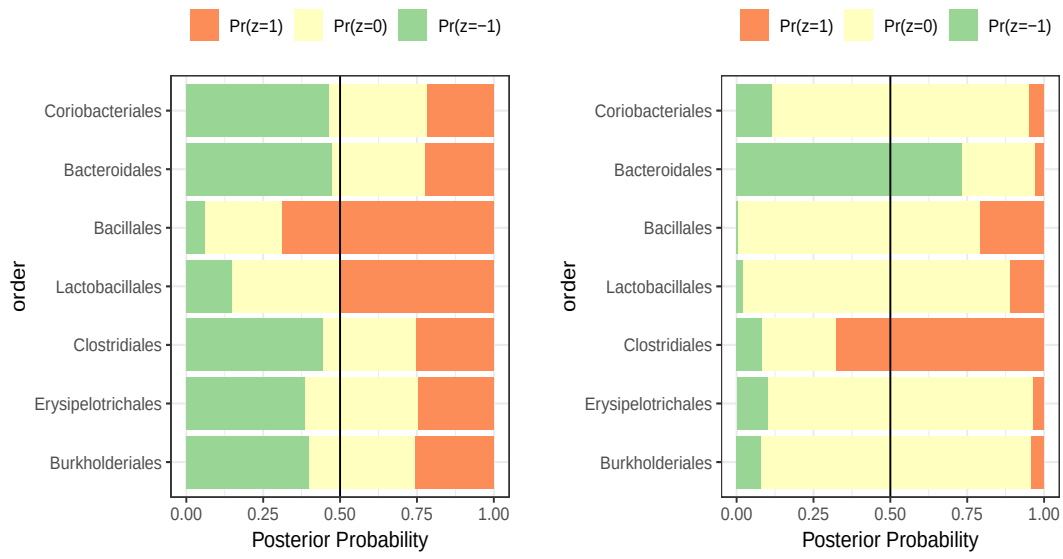
Table 4.4 shows that estimated direct effect of high fat on BMI and the mediation effect of the identified balance and their 95 credible intervals. We observe a strong direct effect of high fat on BMI and also the strong effect of the identified balance on the BMI, but a relatively weak mediating effect of the balance.

Table 4.4: Results of Bayesian balance mediation analysis at phylum and order level, assuming uniform or sparse prior for z. The table shows the estimation of direct, mediation effects on high-fat on BMI and the effect of balance on BMI. based on the predictive posterior distributions.

	Phylum level		Order level	
	Uniform Prior	Sparse Prior	Uniform prior	Sparse prior
High fat(T)	1.90 (1.01,2.10)	2.01 (0.99,2.10)	1.46 (1.02,2.06)	1.86 (1.15,2.11
Balance	0.69 (0,0.92)	0.84 (0.01,0.92)	0.04 (0,0.86)	0.58 (0,0.92)
Mediation	0.17 (-0.76,1.04)	0.19 (-0.32,0.83)	0.03 (-2.12,2.19)	0.19(-1.33,1.50)

4.6.2. Analysis at the order level

Our second analysis is performed at the order level, where we summarize the baterical counts data at the order level, and for each sample, we have read count for each of the 8 different orders, including *Coriobacteriales, Bacteroidales, Bacillales, Lactobacillales, Clostridiales, Erysipelotrichales*, and *Burkholderiales*. In our analysis, we choose *Burkholderiales* as the reference order in the logistic-normal distribution. Figure 4.18 shows the plot of the posterior distributions of the balance configuration of seven bacterial orders in Bayesian balance mediation analysis of the effect of high-fat on BMI. Under the uniform prior, no clear balance can be identified that mediate the effect of high fat on MBI. Under the sparse prior, the plot indicates that the ratio of order of *Costridiales* and order *Bacteroidels* defines the balance that mediates the effect of high fat on BMI. Since *Clostridiales* belong to phylum *Frmicutes* and *Bacteroidales* belongs to phylum *Bacteroide*, the balance identified at the order level agrees with our analysis at the phylum level.



(a) Uniform prior

(b) Sparse prior

Figure 4.18: Plot of the posterior distribution of the balance configuration of 7 bacterial orders in Bayesian balance mediation analysis of the effect of high-fat on BMI.

Table 4.4 shows the estimate of the direct effect of high fat on BMI and the mediation effect by the identified balance and their 95 credible intervals. The estimated direct effect and mediation effect are similar to our analysis at the phylum level. We observe a strong direct effect of high fat on BMI and strong effect of the balance identified on BMI, but a relatively weak mediating effect.

4.7. Discussion

In this work, we have described a general Bayesian balance mediation model that extends the compositional mediation analysis via the balance index introduced in earlier chapters. The balance is able to characterize a microbiome sample with a single scalar and the mediation analysis based on the balance index is advantageous with regard to dimension reduction and model interpretation. We have developed a generative model in a Bayesian framework to jointly model the taxon read counts and an continuous outcome variable via the latent bacterial compositions. In particular, we have developed a two group mediation analysis method where the treatment/exposure is a binary variable in order to estimate the treatment effect that is mediated through perturbing the microbiome composition. We have developed a hybrid Gibbs-Metropolis sampling algorithm to sample all the

unknown quantities from the posterior distributions. The mean vector and variance matrix for the compositions in the treated and untreated groups are updated with the observed count as well as the latent relative abundance, while the balance indicator is updated using the latent relative abundance and the outcomes. The estimates for path coefficients and the indirect mediation effects are calculated for each sampling realization and are used for posterior inference. Simulation studies have shown the effectiveness of the proposed methods in identifying the mediating balance and its mediation effect.

Our proposed balance mediation model is fully parametric, which facilitates sampling from the posterior distributions. The posterior sampling makes statistical inference easier than the traditional mediation analysis methods or other high dimensional compositional mediation analysis, which often rely on bootstrapping. On the other hand, similar to all Bayesian methods, our method can be sensitive to hyperparameters when sample size is small. In the case when we have enough data, the impact of the prior distributions on the inference is minimal. However, when the sample size is small, the posterior distributions can be dominated by the prior distributions of the parameters. As we observed in our real data analysis, when the sample sizes are small, uniform or sparse prior assumption on the balance configuration vector $\mathbf{z}$ can lead to identifying different sets of taxa that define the balance.

The proposed Bayesian mediation analysis is computational intensive. Since the latent relative abundance matrix $\mathbf{P}$ needs to be sampled at each iteration, if n and q are big, it can take a considerable time even to run a couple hundred iterations. In high dimensional settings, we have to run many steps of the posterior sampling in order to reach convergence. The benchmark for all the simulation studies presented above is 3 minutes and 7 seconds for $n = 100, q = 10$ with $3 * 10^4$ total iterations. In our analysis of the real data sets, we have focused our analysis at the phylum or family level, where the number of taxa q is relatively small, which usually takes several minutes. When we apply the methods to the compositional data at the species level, we expect longer running time. For such a high dimensional taxa composition, it makes more biological sense to identify smaller balance by assuming a sparse prior on the balance configuration.

There are several potential extensions to our proposed model. First, in relating the relative abundance to count data, we use a multinomial model. It is possible to model the microbiome count data using some zero inflated modes, due to the sparsity of the data from microbiome studies. However,

if we can aggregate the data into higher levels on the taxonomy tree and remove the taxon that only appears in less than 10% subjects, a multinomial model is often sufficient. Second, although for simplicity, we present our mediation analysis for binary treatment, the model and the methods can be similarly extended for continuous treatment and for including covariates. For binary treatment, there is a one-to-one correspondence between $\boldsymbol{a}$ and $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$. For general continuous factor T_j, we can replace the location parameters $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$ in the distribution of $\boldsymbol{p}$ by $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(T_j - \bar{T})$. This parameterization allows interpretation of $\boldsymbol{\beta}_0$ as the location of log-ratio transformation of the compositions when $T_j = \bar{T}$ and $\boldsymbol{\beta}_1$ as the change in location for a unit change in T_j. In this case, the mediation effect can still be quantified by $bal(a, \boldsymbol{z}) \times b_z$ for a given balance configuration $\boldsymbol{z}$. Similar MCMC algorithm can be developed.

In our real data analysis, the mediation effect of the identified balance is relatively weak compared to the direct effect of vegan diet on all four metabolites. The production of the selected metabolites requires substrate (fiber-rich food) as well as gut microbiota. As the original study pointed out (Wu et al., 2016), there is only a small difference in the composition of gut microbiota between subjects with vegan and omnivore diet, the observed elevated level of the four metabolites in vegan diet is probably due to the availability of fiber-rich food. The changes in composition of microbiota is slow with a shift in diet pattern. A future study with a sufficiently large number of subjects who are on vegan diet for a long time would lead to a thorough understanding of the balance configuration and its role in mediating the effect of vegan diet on each of the metabolite.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTION

5.1. Discussion and conclusions

This thesis work develops novel Bayesian balance regression and Bayesian balance mediation analysis in order to identify that microbial compositional balance that is associated with outcomes or serves as possible mediator between treatment and outcome. Our approaches overcome two major challenges in analysis of microbiome data, high-dimensionality and compositional nature of the data. The balance provides a summary measure of dysbiosis of the microbial community. With the balance index, physicians are able to evaluate a patient's gut microbiota healthy status or his/her risk of getting a particular gastrointestinal disease, or even assess whether a treatment is effective for a specific patient. The simple calculation and easy use of a scalar measure derived from a high dimensional compositional data can potentially facilitate the use of microbiome data in medical practice.

Throughout my thesis work, I have explored how to extract the balance index under different statistical models, which may have direct medical applications. The main theme is to recover the balance index in the regression setting, where the balance composition can be recovered as well as its association with an outcome. The MCMC sampling algorithms for the proposed Bayesian regression and Bayesian mediation analysis enable efficient exploration of the large latent space of balance indicator. To speed up the convergence of Markov chains, the unrelated parameters are integrated out. For binary outcome regression, we use the probit link function because there is simple one-to-one correspondence between the probit model and linear model. By augmenting the data with an unobserved latent continuous variable, a new sampling procedure is introduced where the latent variable is first updated and then the recovery of a balance index is performed with the new value of latent variable. For the proposed Bayesian mediation analysis, balance index and compositional vector are introduced into the linear structure equation model. In this work, we directly model the count data and use the unobserved microbiome composition to associate the bacterial read counts and the outcome.

Like all Bayesian methods, the choice of prior distributions might impact the results if the sample

size is not large enough. However, if there is strong evidence supporting a particular choice of prior distributions, it should definitely help to recover the balance index using additional data. When researchers have no prior knowledge, an uninformative or a flat prior is often chosen, and in such case, the results are largely driven by data. The second drawback of the proposed Bayesian approach is computation cost and time. MCMC sampling requires a lot of computational time even we integrate out the unrelated parameters in the model when sampling the balance index. In recovering the balance index using probit model, an additional step is required to sample the latent variable from a truncated multivariate normal distribution. This step deteriorates the performance of proposed method further. Similarly, in the balance mediation model, the relative abundance vector for all subjects are latent, which requires additional sampling steps and more computational time.

5.2. Future work

The balance presents a simple and yet powerful summary of microbiome composition that can be used as a measure of dysbiosis of a microbial community. In addition to what I have completed so far, the balance index can fit into any statistical models for high dimensional compositional data. For example, instead of a binary outcome, we can incorporate balance index to a generalized linear model and Cox proportional hazards regression model. Similarly we can build balance mediation model for different type of outcome data. Due to the power of Bayesian sampling and posterior inference, we can estimate parameters from a much complex model even without closed-form expressions of the posterior distribution.

Besides possible extensions of the balance regression to various other types of regression problems, another interesting question is to identify the balance in an unsupervised setting. Inspired by the PCA where the first principle component corresponds to the direction that has the largest variance, it is interesting to identify balance that has the largest variance among given samples. Let $\mathbf{x}_i$ be the compositional vector for the ith sample in a given study of n independent samples, we can identify balance indicator vector that maximizes the sample variance of the corresponding balance,

$$\max_{\mathbf{z}} Var\{Bal(\mathbf{x_i}, \mathbf{z}) = \frac{1}{n-1} \max_{\mathbf{z}} \sum_i (Bal(\mathbf{x}_i, \mathbf{z}) - \bar{Bal}(\mathbf{z}))^2,$$

This optimization can be solved using the adaptive hyperbox algorithm developed for high dimensional discrete optimization (Xu, Nelson, and Hong, 2013).

APPENDIX A

ADDITIONAL REAL DATA ANALYSIS RESULTS UNDER DIFFERENT MCMC INITIAL VALUES

A.1. Chapter 2 - additional results on analysis of COMBO and UK twin data sets

A.1.1. Additional results on COMBO data

Table A.1: Posterior probabilities with 5 starting values under uniform prior. Results are from COMBO data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Collinsella	0.16	0.50	0.34	0.15	0.51	0.34	0.15	0.51	0.34	0.17	0.49	0.34	0.16	0.49	0.35
Eggerthella	0.23	0.41	0.36	0.24	0.43	0.34	0.24	0.41	0.35	0.23	0.42	0.35	0.24	0.42	0.34
Bacteroides	0.34	0.32	0.34	0.34	0.31	0.34	0.34	0.31	0.35	0.33	0.32	0.35	0.34	0.31	0.34
Barnesiella	0.19	0.17	0.64	0.19	0.17	0.64	0.19	0.16	0.65	0.20	0.19	0.62	0.19	0.19	0.62
Butyricimonas	0.12	0.41	0.47	0.12	0.43	0.45	0.12	0.41	0.47	0.12	0.42	0.45	0.11	0.42	0.46
Odoribacter	0.06	0.57	0.38	0.05	0.57	0.38	0.05	0.59	0.36	0.07	0.56	0.37	0.06	0.56	0.38
Parabacteroides	0.32	0.26	0.42	0.31	0.24	0.45	0.31	0.25	0.44	0.34	0.24	0.42	0.33	0.25	0.42
Paraprevotella	0.25	0.29	0.47	0.25	0.29	0.47	0.25	0.28	0.47	0.26	0.28	0.45	0.26	0.28	0.46
Prevotella	0.04	0.46	0.49	0.04	0.45	0.51	0.04	0.48	0.48	0.05	0.44	0.51	0.04	0.45	0.51
Alistipes	0.02	0.79	0.19	0.02	0.78	0.20	0.02	0.80	0.18	0.02	0.77	0.21	0.02	0.79	0.19
Gemella	0.32	0.34	0.34	0.31	0.35	0.34	0.30	0.35	0.35	0.31	0.35	0.34	0.32	0.35	0.33
Granulicatella	0.33	0.33	0.34	0.32	0.33	0.35	0.33	0.33	0.34	0.34	0.34	0.32	0.35	0.33	0.32
Lactobacillus	0.25	0.37	0.38	0.28	0.35	0.37	0.26	0.37	0.38	0.29	0.34	0.37	0.26	0.38	0.36
Streptococcus	0.23	0.38	0.39	0.24	0.38	0.38	0.25	0.37	0.38	0.25	0.36	0.39	0.25	0.37	0.38
Clostridium	0.03	0.79	0.18	0.03	0.77	0.20	0.03	0.80	0.17	0.03	0.78	0.18	0.03	0.79	0.18
Anaerofustis	0.30	0.35	0.35	0.31	0.33	0.36	0.31	0.35	0.34	0.32	0.33	0.35	0.31	0.36	0.33
Eubacterium	0.20	0.32	0.48	0.21	0.31	0.48	0.20	0.32	0.48	0.21	0.32	0.46	0.21	0.33	0.47
Anaerovorax	0.16	0.51	0.33	0.14	0.51	0.35	0.14	0.51	0.35	0.14	0.51	0.35	0.14	0.51	0.35
Mogibacterium	0.44	0.23	0.33	0.46	0.22	0.32	0.44	0.24	0.32	0.45	0.24	0.31	0.43	0.25	0.32
Blautia	0.31	0.31	0.39	0.32	0.33	0.35	0.32	0.32	0.36	0.32	0.31	0.37	0.31	0.33	0.36
Coprococcus	0.21	0.37	0.43	0.19	0.38	0.43	0.20	0.37	0.43	0.23	0.36	0.41	0.21	0.36	0.42
Dorea	0.66	0.08	0.25	0.67	0.09	0.25	0.69	0.08	0.23	0.69	0.08	0.23	0.68	0.09	0.24
Roseburia	0.10	0.56	0.35	0.09	0.57	0.34	0.10	0.57	0.33	0.11	0.56	0.33	0.10	0.55	0.35
Anaerofilum	0.23	0.42	0.34	0.23	0.43	0.34	0.22	0.43	0.35	0.24	0.43	0.34	0.23	0.42	0.35
Anaerotruncus	0.17	0.42	0.40	0.18	0.44	0.38	0.17	0.45	0.38	0.19	0.42	0.39	0.17	0.43	0.41
Butyricicoccus	0.27	0.35	0.38	0.26	0.34	0.39	0.27	0.34	0.38	0.28	0.32	0.40	0.29	0.34	0.37
Faecalibacterium	0.48	0.16	0.36	0.50	0.15	0.35	0.50	0.16	0.34	0.49	0.16	0.35	0.46	0.18	0.36
Oscillibacter	0.04	0.73	0.23	0.04	0.72	0.24	0.03	0.73	0.24	0.04	0.74	0.22	0.03	0.72	0.24

Table A.1: Posterior probabilities with 5 starting values under uniform prior. Results are from COMBO data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Ruminococcus	0.58	0.08	0.34	0.60	0.08	0.32	0.62	0.06	0.32	0.59	0.08	0.33	0.61	0.08	0.31
Subdoligranulum	0.35	0.21	0.44	0.34	0.21	0.44	0.35	0.21	0.44	0.33	0.21	0.46	0.33	0.22	0.44
Acidaminococcus	0.95	0.00	0.04	0.95	0.01	0.05	0.95	0.01	0.04	0.94	0.01	0.06	0.94	0.01	0.06
Allisonella	0.78	0.05	0.16	0.78	0.06	0.16	0.78	0.05	0.17	0.77	0.06	0.18	0.76	0.06	0.18
Dialister	0.03	0.62	0.35	0.03	0.61	0.36	0.03	0.63	0.34	0.04	0.63	0.33	0.04	0.64	0.32
Megamonas	0.06	0.69	0.25	0.05	0.69	0.26	0.06	0.66	0.27	0.07	0.67	0.26	0.06	0.69	0.25
Megasphaera	0.79	0.04	0.17	0.77	0.04	0.19	0.80	0.04	0.16	0.78	0.05	0.17	0.77	0.05	0.17
P. faecium	0.06	0.43	0.51	0.07	0.42	0.51	0.06	0.43	0.50	0.06	0.44	0.50	0.06	0.43	0.51
Veillonella	0.13	0.53	0.34	0.12	0.53	0.35	0.12	0.54	0.34	0.13	0.53	0.34	0.13	0.54	0.33
Catenibacterium	0.74	0.06	0.20	0.73	0.06	0.21	0.73	0.06	0.22	0.71	0.06	0.23	0.72	0.07	0.21
Coprobacillus	0.04	0.66	0.30	0.04	0.67	0.29	0.05	0.66	0.30	0.04	0.66	0.30	0.05	0.66	0.29
Holdemania	0.25	0.35	0.40	0.26	0.36	0.37	0.27	0.33	0.40	0.27	0.35	0.38	0.27	0.34	0.39
Solobacterium	0.26	0.40	0.34	0.25	0.40	0.35	0.26	0.39	0.35	0.28	0.37	0.36	0.26	0.39	0.35
Turicibacter	0.38	0.28	0.34	0.36	0.25	0.38	0.37	0.28	0.36	0.37	0.28	0.35	0.37	0.28	0.35
Parasutterella	0.14	0.34	0.51	0.13	0.34	0.53	0.14	0.36	0.50	0.17	0.35	0.48	0.14	0.36	0.50
Sutterella	0.21	0.30	0.49	0.21	0.30	0.49	0.20	0.30	0.50	0.23	0.29	0.48	0.21	0.32	0.47
Oxalobacter	0.17	0.48	0.35	0.19	0.48	0.33	0.18	0.50	0.32	0.19	0.47	0.34	0.18	0.47	0.35

Table A.2: Posterior probabilities with 5 starting values under sparse prior. Results are from COMBO data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Collinsella	0.02	0.18	0.80	0.01	0.19	0.79	0.01	0.20	0.78	0.01	0.21	0.78	0.01	0.19	0.79
Eggerthella	0.03	0.18	0.79	0.03	0.20	0.77	0.03	0.18	0.79	0.03	0.19	0.79	0.02	0.20	0.78
Bacteroides	0.08	0.06	0.87	0.07	0.06	0.87	0.07	0.06	0.88	0.07	0.07	0.86	0.07	0.07	0.86
Barnesiella	0.00	0.01	0.99	0.00	0.01	0.99	0.01	0.01	0.99	0.00	0.01	0.99	0.00	0.01	0.99
Butyricimonas	0.00	0.03	0.96	0.01	0.03	0.96	0.00	0.03	0.96	0.00	0.03	0.97	0.01	0.03	0.96
Odoribacter	0.00	0.07	0.93	0.01	0.07	0.93	0.01	0.07	0.92	0.00	0.08	0.92	0.00	0.05	0.94

Table A.2: Posterior probabilities with 5 starting values under sparse prior. Results are from COMBO data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Parabacteroides	0.02	0.02	0.95	0.02	0.03	0.95	0.02	0.03	0.95	0.02	0.03	0.95	0.02	0.03	0.95
Paraprevotella	0.01	0.03	0.97	0.01	0.03	0.97	0.01	0.03	0.96	0.01	0.03	0.96	0.01	0.03	0.97
Prevotella	0.00	0.03	0.97	0.00	0.03	0.97	0.00	0.03	0.97	0.00	0.03	0.97	0.00	0.03	0.97
Alistipes	0.00	0.64	0.36	0.00	0.62	0.38	0.00	0.64	0.36	0.00	0.61	0.39	0.00	0.65	0.35
Gemella	0.07	0.12	0.81	0.07	0.11	0.81	0.08	0.12	0.81	0.07	0.12	0.81	0.07	0.12	0.81
Granulicatella	0.08	0.10	0.82	0.07	0.10	0.83	0.07	0.09	0.84	0.07	0.10	0.83	0.07	0.10	0.83
Lactobacillus	0.03	0.08	0.89	0.04	0.09	0.88	0.04	0.10	0.86	0.04	0.10	0.87	0.03	0.10	0.86
Streptococcus	0.02	0.06	0.92	0.02	0.07	0.91	0.02	0.07	0.91	0.02	0.07	0.92	0.02	0.06	0.92
Clostridium	0.00	0.81	0.19	0.00	0.78	0.22	0.00	0.79	0.21	0.00	0.76	0.24	0.00	0.78	0.22
Anaerofustis	0.07	0.12	0.81	0.07	0.13	0.80	0.07	0.12	0.81	0.07	0.13	0.80	0.07	0.12	0.81
Eubacterium	0.01	0.04	0.96	0.01	0.03	0.96	0.01	0.03	0.96	0.01	0.04	0.96	0.01	0.04	0.96
Anaerovorax	0.02	0.15	0.83	0.01	0.15	0.84	0.02	0.15	0.83	0.02	0.14	0.84	0.02	0.16	0.82
Mogibacterium	0.13	0.05	0.81	0.15	0.04	0.81	0.13	0.05	0.82	0.13	0.05	0.81	0.14	0.05	0.81
Blautia	0.04	0.06	0.90	0.05	0.06	0.89	0.04	0.06	0.89	0.04	0.07	0.89	0.04	0.06	0.90
Coprococcus	0.04	0.02	0.94	0.04	0.03	0.93	0.03	0.02	0.94	0.03	0.02	0.94	0.03	0.02	0.94
Dorea	0.30	0.01	0.69	0.35	0.00	0.65	0.32	0.01	0.68	0.33	0.01	0.67	0.30	0.01	0.69
Roseburia	0.00	0.10	0.89	0.00	0.10	0.89	0.01	0.10	0.89	0.01	0.10	0.89	0.01	0.10	0.89
Anaerofilum	0.05	0.17	0.78	0.04	0.16	0.80	0.05	0.15	0.80	0.05	0.16	0.80	0.05	0.14	0.81
Anaerotruncus	0.02	0.05	0.93	0.01	0.05	0.93	0.02	0.05	0.93	0.01	0.05	0.94	0.01	0.05	0.94
Butyricicoccus	0.03	0.06	0.91	0.03	0.06	0.91	0.03	0.05	0.92	0.03	0.06	0.91	0.03	0.06	0.91
Faecalibacterium	0.11	0.01	0.87	0.11	0.01	0.88	0.10	0.01	0.89	0.11	0.01	0.88	0.10	0.01	0.89
Oscillibacter	0.00	0.26	0.74	0.00	0.27	0.73	0.00	0.26	0.73	0.00	0.28	0.72	0.00	0.24	0.75
Ruminococcus	0.07	0.00	0.93	0.05	0.00	0.94	0.06	0.00	0.94	0.06	0.00	0.93	0.05	0.00	0.95
Subdoligranulum	0.08	0.01	0.92	0.06	0.01	0.93	0.06	0.01	0.93	0.06	0.01	0.92	0.07	0.01	0.92
Acidaminococcus	0.93	0.00	0.07	0.91	0.00	0.09	0.91	0.00	0.09	0.90	0.00	0.10	0.88	0.00	0.12
Allisonella	0.62	0.01	0.37	0.64	0.00	0.35	0.63	0.01	0.37	0.65	0.01	0.34	0.66	0.00	0.34
Dialister	0.00	0.04	0.96	0.00	0.05	0.95	0.00	0.05	0.95	0.00	0.04	0.96	0.00	0.04	0.96
Megamonas	0.00	0.55	0.45	0.00	0.48	0.52	0.00	0.49	0.51	0.00	0.49	0.51	0.00	0.49	0.51
Megasphaera	0.24	0.01	0.76	0.24	0.01	0.75	0.22	0.01	0.77	0.20	0.01	0.79	0.25	0.01	0.74

Table A.2: Posterior probabilities with 5 starting values under sparse prior. Results are from COMBO data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
P. faecium	0.00	0.02	0.98	0.00	0.02	0.98	0.00	0.02	0.98	0.00	0.02	0.98	0.00	0.02	0.98
Veillonella	0.01	0.18	0.81	0.01	0.18	0.81	0.01	0.17	0.82	0.00	0.16	0.83	0.01	0.17	0.82
Catenibacterium	0.57	0.01	0.42	0.47	0.00	0.53	0.50	0.00	0.49	0.49	0.00	0.50	0.47	0.01	0.52
Coprobacillus	0.00	0.15	0.85	0.00	0.17	0.83	0.00	0.17	0.83	0.00	0.19	0.81	0.00	0.17	0.83
Holdemania	0.02	0.07	0.91	0.01	0.08	0.90	0.02	0.08	0.91	0.01	0.07	0.91	0.02	0.08	0.91
Solobacterium	0.04	0.15	0.82	0.03	0.17	0.80	0.03	0.16	0.80	0.03	0.16	0.81	0.04	0.17	0.79
Turicibacter	0.16	0.03	0.81	0.14	0.04	0.82	0.12	0.03	0.85	0.12	0.04	0.84	0.14	0.03	0.83
Parasutterella	0.01	0.03	0.97	0.01	0.02	0.97	0.00	0.02	0.97	0.00	0.02	0.98	0.00	0.02	0.98
Sutterella	0.01	0.03	0.96	0.01	0.02	0.97	0.00	0.03	0.97	0.00	0.03	0.97	0.00	0.03	0.97
Oxalobacter	0.02	0.16	0.82	0.02	0.18	0.81	0.02	0.16	0.81	0.02	0.18	0.80	0.02	0.16	0.82

A.1.2. Additional results on UK twin data

Table A.3: Posterior probabilities with 5 starting values under uniform prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Acidaminococcus	0.00	0.58	0.42	0.00	0.61	0.39	0.00	0.60	0.39	0.00	0.63	0.37	0.00	0.61	0.39
Actinomyces	0.90	0.00	0.10	0.88	0.00	0.12	0.89	0.00	0.11	0.89	0.00	0.11	0.90	0.00	0.10
Adlercreutzia	0.33	0.07	0.60	0.36	0.07	0.57	0.34	0.08	0.58	0.33	0.07	0.60	0.36	0.07	0.57
Aggregatibacter	0.03	0.58	0.39	0.03	0.59	0.39	0.02	0.62	0.36	0.03	0.57	0.40	0.02	0.62	0.37
Akkermansia	0.36	0.00	0.64	0.35	0.00	0.65	0.31	0.00	0.69	0.37	0.00	0.63	0.32	0.00	0.68
Alistipes	0.42	0.06	0.52	0.42	0.08	0.49	0.40	0.08	0.51	0.42	0.08	0.50	0.44	0.07	0.49
Anaerofustis	0.28	0.13	0.59	0.29	0.13	0.58	0.27	0.15	0.58	0.29	0.14	0.57	0.27	0.15	0.59
Anaerostipes	0.00	0.94	0.06	0.00	0.94	0.06	0.00	0.92	0.08	0.00	0.92	0.08	0.00	0.92	0.08
Anaerotruncus	0.31	0.11	0.58	0.30	0.13	0.56	0.30	0.11	0.59	0.30	0.11	0.59	0.31	0.12	0.57
Anaerovorax	0.67	0.03	0.30	0.67	0.03	0.30	0.65	0.03	0.32	0.65	0.04	0.32	0.62	0.03	0.35
Bacteroides	0.01	0.87	0.12	0.01	0.87	0.13	0.01	0.85	0.14	0.01	0.87	0.13	0.01	0.85	0.14

Table A.3: Posterior probabilities with 5 starting values under uniform prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Bifidobacterium	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Bilophila	0.25	0.03	0.73	0.22	0.03	0.75	0.25	0.03	0.73	0.23	0.02	0.74	0.22	0.03	0.75
Blautia	0.01	0.84	0.14	0.01	0.86	0.13	0.01	0.85	0.14	0.01	0.86	0.13	0.01	0.86	0.13
Bulleidia	0.65	0.02	0.33	0.69	0.02	0.29	0.67	0.01	0.32	0.65	0.02	0.33	0.69	0.02	0.29
Butyricimonas	0.10	0.03	0.87	0.11	0.02	0.87	0.10	0.02	0.87	0.10	0.02	0.88	0.10	0.02	0.87
Campylobacter	0.25	0.21	0.54	0.26	0.23	0.51	0.27	0.21	0.52	0.25	0.24	0.51	0.25	0.21	0.53
Catenibacterium	0.26	0.01	0.73	0.26	0.01	0.73	0.27	0.01	0.72	0.26	0.01	0.73	0.27	0.01	0.72
cc_115	0.22	0.04	0.74	0.22	0.04	0.74	0.22	0.04	0.74	0.23	0.05	0.72	0.23	0.04	0.73
Christensenella	0.01	0.86	0.14	0.00	0.83	0.16	0.01	0.81	0.19	0.00	0.83	0.17	0.00	0.82	0.17
Citrobacter	0.45	0.01	0.54	0.47	0.01	0.52	0.47	0.01	0.52	0.49	0.01	0.50	0.47	0.01	0.52
Clostridium	0.65	0.00	0.35	0.67	0.00	0.33	0.65	0.00	0.35	0.68	0.00	0.31	0.63	0.00	0.36
Collinsella	0.16	0.00	0.84	0.16	0.01	0.83	0.15	0.01	0.84	0.17	0.00	0.83	0.16	0.01	0.83
Coprobacillus	0.75	0.00	0.24	0.76	0.00	0.24	0.74	0.00	0.25	0.77	0.00	0.23	0.75	0.00	0.25
Coprococcus	0.11	0.52	0.37	0.11	0.48	0.41	0.11	0.49	0.40	0.10	0.53	0.37	0.10	0.50	0.40
Corynebacterium	0.11	0.49	0.40	0.12	0.47	0.41	0.11	0.48	0.41	0.13	0.44	0.43	0.10	0.49	0.41
Dehalobacterium	0.14	0.23	0.63	0.15	0.25	0.60	0.14	0.24	0.61	0.14	0.27	0.59	0.14	0.24	0.62
Desulfovibrio	0.27	0.00	0.73	0.29	0.00	0.71	0.31	0.00	0.69	0.29	0.00	0.71	0.27	0.00	0.73
Dialister	0.00	1.00	0.00	0.00	0.99	0.01	0.00	1.00	0.00	0.00	0.99	0.01	0.00	1.00	0.00
Dorea	0.37	0.15	0.48	0.34	0.17	0.50	0.35	0.17	0.48	0.36	0.15	0.49	0.35	0.15	0.50
Eggerthella	0.21	0.16	0.63	0.23	0.19	0.59	0.22	0.18	0.60	0.21	0.18	0.61	0.20	0.20	0.60
Enterococcus	0.25	0.07	0.67	0.26	0.08	0.65	0.26	0.08	0.66	0.25	0.08	0.66	0.26	0.08	0.66
Faecalibacterium	0.01	0.88	0.11	0.01	0.88	0.11	0.00	0.88	0.11	0.00	0.90	0.10	0.00	0.89	0.11
Fusobacterium	0.20	0.15	0.65	0.23	0.17	0.60	0.22	0.15	0.63	0.21	0.16	0.63	0.22	0.15	0.64
Haemophilus	0.09	0.06	0.85	0.10	0.09	0.81	0.12	0.07	0.81	0.10	0.07	0.83	0.10	0.06	0.84
Holdemania	0.59	0.02	0.39	0.61	0.02	0.37	0.62	0.02	0.36	0.62	0.01	0.37	0.63	0.02	0.35
Lachnobacterium	0.21	0.07	0.71	0.22	0.08	0.71	0.21	0.08	0.71	0.23	0.08	0.69	0.20	0.07	0.73
Lachnospira	0.76	0.00	0.24	0.76	0.01	0.23	0.73	0.00	0.27	0.75	0.00	0.25	0.74	0.01	0.26
Lactobacillus	0.98	0.00	0.02	0.98	0.00	0.02	0.97	0.00	0.03	0.98	0.00	0.02	0.98	0.00	0.02
Lactococcus	0.10	0.39	0.51	0.11	0.40	0.49	0.12	0.39	0.49	0.11	0.41	0.48	0.12	0.41	0.47

Table A.3: Posterior probabilities with 5 starting values under uniform prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Megasphaera	0.17	0.09	0.73	0.17	0.11	0.72	0.16	0.11	0.73	0.18	0.10	0.72	0.16	0.12	0.72
Methanobrevibacter	0.50	0.00	0.50	0.51	0.00	0.49	0.52	0.00	0.48	0.50	0.00	0.50	0.53	0.00	0.47
Odoribacter	0.65	0.00	0.35	0.64	0.00	0.36	0.61	0.00	0.39	0.63	0.00	0.37	0.58	0.00	0.41
Oscillospira	0.29	0.21	0.50	0.30	0.19	0.51	0.31	0.22	0.48	0.31	0.21	0.49	0.31	0.21	0.48
Oxalobacter	0.05	0.14	0.81	0.05	0.16	0.78	0.06	0.16	0.78	0.05	0.15	0.80	0.05	0.15	0.80
Parabacteroides	0.00	0.65	0.34	0.00	0.68	0.31	0.00	0.68	0.31	0.00	0.68	0.31	0.00	0.71	0.29
Paraprevotella	0.02	0.01	0.98	0.02	0.02	0.96	0.02	0.02	0.96	0.03	0.01	0.96	0.02	0.02	0.97
P. faecium	0.02	0.00	0.98	0.03	0.00	0.97	0.02	0.00	0.98	0.02	0.00	0.98	0.03	0.00	0.96
Porphyromonas	0.43	0.04	0.53	0.44	0.04	0.52	0.42	0.05	0.53	0.43	0.04	0.53	0.42	0.05	0.53
Prevotella	0.01	0.01	0.99	0.02	0.01	0.97	0.01	0.00	0.99	0.01	0.01	0.98	0.01	0.00	0.98
Ralstonia	0.12	0.32	0.55	0.13	0.37	0.51	0.12	0.36	0.52	0.12	0.35	0.53	0.13	0.34	0.53
rc4-4	0.41	0.01	0.58	0.39	0.01	0.60	0.42	0.01	0.57	0.38	0.01	0.62	0.40	0.01	0.59
Roseburia	0.07	0.38	0.55	0.08	0.38	0.54	0.05	0.43	0.52	0.07	0.38	0.54	0.05	0.41	0.55
Rothia	0.34	0.10	0.56	0.36	0.10	0.54	0.34	0.12	0.55	0.35	0.10	0.56	0.33	0.12	0.55
Ruminococcus	0.50	0.08	0.42	0.53	0.08	0.39	0.52	0.08	0.40	0.54	0.07	0.40	0.52	0.07	0.41
Serratia	0.41	0.03	0.56	0.44	0.03	0.53	0.43	0.03	0.54	0.42	0.04	0.55	0.41	0.03	0.56
Slackia	0.52	0.01	0.48	0.52	0.01	0.47	0.56	0.01	0.43	0.52	0.01	0.47	0.55	0.01	0.45
SMB53	0.09	0.44	0.47	0.10	0.43	0.47	0.08	0.42	0.49	0.08	0.47	0.45	0.09	0.45	0.46
Streptococcus	0.98	0.00	0.02	0.98	0.00	0.02	0.99	0.00	0.01	0.98	0.00	0.02	0.99	0.00	0.01
Sutterella	0.00	0.74	0.26	0.00	0.72	0.27	0.00	0.70	0.30	0.00	0.71	0.28	0.00	0.67	0.33
Trabulsiella	0.95	0.00	0.05	0.95	0.00	0.05	0.93	0.00	0.07	0.94	0.00	0.06	0.94	0.00	0.06
Turicibacter	0.82	0.00	0.18	0.80	0.00	0.20	0.84	0.00	0.16	0.83	0.00	0.17	0.81	0.00	0.19
Veillonella	1.00	0.00	0.00	1.00	0.00	0.00	0.99	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.01
WAL_1855D	0.00	0.94	0.06	0.00	0.94	0.05	0.00	0.94	0.06	0.00	0.95	0.05	0.00	0.93	0.07
Eubacterium	0.00	0.34	0.66	0.00	0.36	0.64	0.00	0.41	0.59	0.00	0.34	0.65	0.00	0.39	0.61

Table A.4: Posterior probabilities with 5 starting values under sparse prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Acidaminococcus	0.00	0.08	0.92	0.00	0.09	0.91	0.00	0.07	0.93	0.00	0.08	0.92	0.00	0.08	0.92
Actinomyces	0.70	0.00	0.30	0.68	0.00	0.32	0.75	0.00	0.24	0.69	0.00	0.31	0.73	0.00	0.27
Adlercreutzia	0.10	0.00	0.90	0.09	0.00	0.90	0.11	0.00	0.89	0.11	0.00	0.89	0.11	0.00	0.88
Aggregatibacter	0.00	0.18	0.82	0.00	0.17	0.83	0.00	0.16	0.84	0.00	0.19	0.81	0.00	0.16	0.84
Akkermansia	0.03	0.00	0.97	0.03	0.00	0.97	0.03	0.00	0.97	0.02	0.00	0.98	0.03	0.00	0.97
Alistipes	0.10	0.01	0.89	0.09	0.01	0.91	0.11	0.00	0.89	0.09	0.01	0.90	0.11	0.01	0.88
Anaerofustis	0.11	0.00	0.88	0.11	0.01	0.88	0.13	0.00	0.87	0.12	0.00	0.88	0.13	0.01	0.87
Anaerostipes	0.00	0.47	0.53	0.00	0.49	0.51	0.00	0.44	0.56	0.00	0.47	0.53	0.00	0.45	0.55
Anaerotruncus	0.07	0.01	0.92	0.09	0.00	0.91	0.08	0.00	0.92	0.07	0.00	0.92	0.08	0.00	0.92
Anaerovorax	0.46	0.00	0.54	0.43	0.00	0.57	0.45	0.00	0.55	0.49	0.00	0.51	0.46	0.00	0.54
Bacteroides	0.00	0.86	0.14	0.00	0.87	0.13	0.00	0.90	0.10	0.00	0.86	0.13	0.00	0.88	0.12
Bifidobacterium	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Bilophila	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.00	0.00	0.99
Blautia	0.00	0.77	0.23	0.00	0.80	0.20	0.00	0.84	0.16	0.00	0.85	0.15	0.00	0.82	0.18
Bulleidia	0.30	0.00	0.70	0.33	0.00	0.67	0.30	0.00	0.70	0.32	0.00	0.68	0.32	0.00	0.68
Butyricimonas	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.00	0.00	1.00	0.01	0.00	0.99
Campylobacter	0.05	0.03	0.92	0.06	0.03	0.91	0.06	0.02	0.92	0.05	0.02	0.92	0.05	0.02	0.92
Catenibacterium	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.01	0.00	0.99
cc_115	0.05	0.00	0.95	0.07	0.00	0.93	0.05	0.00	0.95	0.08	0.00	0.92	0.06	0.00	0.94
Christensenella	0.00	0.22	0.77	0.00	0.23	0.76	0.00	0.15	0.85	0.00	0.19	0.81	0.00	0.19	0.81
Citrobacter	0.05	0.00	0.95	0.05	0.00	0.95	0.04	0.00	0.96	0.04	0.00	0.96	0.06	0.00	0.94
Clostridium	0.41	0.00	0.59	0.40	0.00	0.60	0.35	0.00	0.65	0.43	0.00	0.57	0.40	0.00	0.60
Collinsella	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99
Coprobacillus	0.54	0.00	0.46	0.52	0.00	0.48	0.51	0.00	0.49	0.51	0.00	0.49	0.53	0.00	0.47
Coprococcus	0.03	0.11	0.86	0.04	0.10	0.86	0.05	0.09	0.86	0.03	0.08	0.89	0.04	0.09	0.87
Corynebacterium	0.02	0.12	0.86	0.02	0.13	0.85	0.02	0.10	0.88	0.02	0.12	0.86	0.02	0.10	0.88
Dehalobacterium	0.12	0.00	0.88	0.11	0.00	0.88	0.11	0.00	0.89	0.10	0.00	0.90	0.11	0.00	0.89
Desulfovibrio	0.00	0.00	1.00	0.01	0.00	0.99	0.00	0.00	1.00	0.01	0.00	0.99	0.01	0.00	0.99
Dialister	0.00	0.48	0.52	0.00	0.56	0.44	0.00	0.41	0.59	0.00	0.54	0.46	0.00	0.44	0.56

Table A.4: Posterior probabilities with 5 starting values under sparse prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Dorea	0.08	0.03	0.88	0.08	0.04	0.88	0.08	0.03	0.89	0.09	0.03	0.88	0.09	0.03	0.89
Eggerthella	0.03	0.03	0.94	0.02	0.03	0.95	0.02	0.02	0.96	0.02	0.02	0.95	0.02	0.02	0.96
Enterococcus	0.02	0.00	0.97	0.02	0.01	0.97	0.03	0.01	0.97	0.02	0.01	0.98	0.02	0.00	0.97
Faecalibacterium	0.00	0.77	0.23	0.00	0.77	0.23	0.00	0.76	0.24	0.00	0.80	0.20	0.00	0.80	0.20
Fusobacterium	0.02	0.02	0.97	0.01	0.02	0.97	0.01	0.01	0.97	0.01	0.01	0.98	0.01	0.01	0.97
Haemophilus	0.00	0.02	0.98	0.00	0.02	0.98	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.01	0.98
Holdemania	0.16	0.00	0.84	0.18	0.00	0.82	0.11	0.00	0.89	0.18	0.00	0.82	0.14	0.00	0.86
Lachnobacterium	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	0.99
Lachnospira	0.44	0.00	0.56	0.40	0.00	0.60	0.43	0.00	0.57	0.39	0.00	0.61	0.44	0.00	0.56
Lactobacillus	0.77	0.00	0.23	0.78	0.00	0.22	0.74	0.00	0.26	0.77	0.00	0.23	0.76	0.00	0.24
Lactococcus	0.01	0.06	0.93	0.01	0.06	0.93	0.01	0.05	0.94	0.01	0.06	0.93	0.01	0.06	0.93
Megasphaera	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.00	0.99	0.00	0.01	0.99	0.00	0.01	0.99
Methanobrevibacter	0.20	0.00	0.80	0.20	0.00	0.80	0.22	0.00	0.78	0.16	0.00	0.84	0.19	0.00	0.81
Odoribacter	0.11	0.00	0.89	0.11	0.00	0.89	0.08	0.00	0.92	0.11	0.00	0.89	0.08	0.00	0.92
Oscillospira	0.08	0.02	0.90	0.10	0.01	0.89	0.09	0.01	0.90	0.09	0.02	0.89	0.08	0.02	0.90
Oxalobacter	0.00	0.00	1.00	0.00	0.00	0.99	0.00	0.00	1.00	0.00	0.00	1.00	0.01	0.00	0.99
Parabacteroides	0.00	0.07	0.93	0.00	0.07	0.93	0.00	0.04	0.96	0.00	0.06	0.94	0.00	0.06	0.94
Paraprevotella	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
P. faecium	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
Porphyromonas	0.21	0.00	0.79	0.22	0.00	0.78	0.22	0.00	0.78	0.19	0.00	0.81	0.21	0.00	0.79
Prevotella	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
Ralstonia	0.01	0.05	0.94	0.01	0.05	0.94	0.01	0.04	0.95	0.01	0.05	0.94	0.01	0.05	0.94
rc4-4	0.02	0.00	0.98	0.02	0.00	0.98	0.02	0.00	0.98	0.03	0.00	0.97	0.02	0.00	0.98
Roseburia	0.00	0.16	0.83	0.00	0.15	0.85	0.00	0.10	0.89	0.00	0.13	0.86	0.00	0.13	0.87
Rothia	0.03	0.03	0.93	0.03	0.04	0.92	0.03	0.03	0.94	0.02	0.04	0.94	0.04	0.04	0.93
Ruminococcus	0.28	0.00	0.71	0.29	0.01	0.70	0.33	0.00	0.66	0.28	0.01	0.72	0.29	0.01	0.71
Serratia	0.04	0.00	0.96	0.03	0.00	0.97	0.03	0.00	0.96	0.03	0.00	0.96	0.04	0.00	0.96
Slackia	0.12	0.00	0.88	0.11	0.00	0.89	0.12	0.00	0.88	0.11	0.00	0.89	0.12	0.00	0.88
SMB53	0.01	0.05	0.94	0.01	0.03	0.96	0.02	0.03	0.95	0.01	0.05	0.93	0.01	0.04	0.95

Table A.4: Posterior probabilities with 5 starting values under sparse prior. Results are from UK twin data at the genus level.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Streptococcus	0.85	0.00	0.15	0.87	0.00	0.13	0.87	0.00	0.13	0.86	0.00	0.14	0.86	0.00	0.14
Sutterella	0.00	0.13	0.87	0.00	0.15	0.85	0.00	0.12	0.88	0.00	0.14	0.86	0.00	0.09	0.91
Trabulsiella	0.92	0.00	0.08	0.93	0.00	0.07	0.94	0.00	0.06	0.94	0.00	0.06	0.92	0.00	0.08
Turicibacter	0.21	0.00	0.79	0.19	0.00	0.81	0.15	0.00	0.85	0.20	0.00	0.80	0.17	0.00	0.83
Veillonella	0.97	0.00	0.03	0.98	0.00	0.02	0.95	0.00	0.05	0.97	0.00	0.03	0.95	0.00	0.05
WAL_1855D	0.00	0.87	0.13	0.00	0.88	0.12	0.00	0.90	0.10	0.00	0.89	0.11	0.00	0.90	0.10
Eubacterium	0.00	0.01	0.99	0.00	0.01	0.99	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00

A.2. Chapter 3 - additional results on analysis of IBD data set

Additional results on IBD data

Table A.5: Posterior probabilities with 5 starting points. Results are from genera level IBD data with uniform prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Prevotella	0.01	0.90	0.09	0.01	0.91	0.08	0.01	0.90	0.09	0.01	0.91	0.08	0.01	0.90	0.09
Faecalibacterium	0.38	0.18	0.45	0.38	0.17	0.45	0.39	0.18	0.44	0.38	0.17	0.45	0.38	0.18	0.44
Eubacterium	0.15	0.45	0.41	0.16	0.43	0.41	0.15	0.44	0.41	0.16	0.42	0.41	0.16	0.43	0.41
Alistipes	0.17	0.39	0.44	0.17	0.38	0.45	0.16	0.39	0.45	0.16	0.39	0.44	0.17	0.39	0.44
Bacteroides	0.33	0.25	0.42	0.33	0.23	0.44	0.35	0.23	0.43	0.33	0.23	0.43	0.31	0.24	0.44
Odoribacter	0.07	0.58	0.35	0.08	0.57	0.36	0.07	0.57	0.36	0.08	0.56	0.35	0.08	0.55	0.38
Dorea	0.36	0.21	0.43	0.35	0.21	0.44	0.33	0.22	0.45	0.37	0.22	0.42	0.34	0.22	0.44
Ruminococcus	0.21	0.37	0.42	0.20	0.38	0.42	0.20	0.37	0.43	0.21	0.37	0.42	0.20	0.37	0.43
Parabacteroides	0.20	0.34	0.46	0.19	0.36	0.45	0.21	0.34	0.45	0.20	0.36	0.44	0.20	0.36	0.44
Akkermansia	0.10	0.52	0.38	0.10	0.53	0.38	0.09	0.52	0.39	0.10	0.52	0.38	0.09	0.53	0.38
Roseburia	0.07	0.58	0.35	0.08	0.57	0.35	0.08	0.57	0.35	0.08	0.57	0.35	0.07	0.57	0.36
Bilophila	0.29	0.25	0.46	0.29	0.25	0.46	0.29	0.24	0.48	0.29	0.24	0.47	0.28	0.26	0.46
Streptococcus	0.29	0.25	0.46	0.30	0.25	0.45	0.29	0.24	0.46	0.31	0.25	0.44	0.30	0.24	0.46
Coprococcus	0.24	0.31	0.45	0.23	0.32	0.45	0.24	0.31	0.45	0.24	0.31	0.45	0.25	0.30	0.45
Collinsella	0.18	0.37	0.45	0.17	0.39	0.44	0.17	0.38	0.44	0.19	0.37	0.44	0.17	0.36	0.46
Clostridium	0.27	0.27	0.45	0.28	0.28	0.43	0.27	0.27	0.47	0.28	0.27	0.45	0.28	0.27	0.45
Sutterella	0.44	0.11	0.45	0.45	0.11	0.45	0.46	0.10	0.44	0.45	0.09	0.45	0.46	0.10	0.44
Bifidobacterium	0.16	0.38	0.47	0.15	0.40	0.45	0.16	0.40	0.45	0.17	0.38	0.45	0.15	0.39	0.46
Anaerotruncus	0.26	0.30	0.43	0.25	0.31	0.44	0.26	0.29	0.45	0.26	0.30	0.44	0.26	0.31	0.43
Lactococcus	0.34	0.20	0.46	0.34	0.20	0.46	0.34	0.20	0.46	0.35	0.20	0.46	0.32	0.21	0.47
Gordonibacter	0.47	0.13	0.40	0.47	0.14	0.39	0.48	0.14	0.38	0.47	0.13	0.39	0.48	0.14	0.38
Eggerthella	0.58	0.08	0.34	0.60	0.07	0.33	0.58	0.07	0.34	0.58	0.07	0.35	0.59	0.07	0.34
Holdemania	0.34	0.21	0.45	0.34	0.23	0.43	0.32	0.22	0.46	0.33	0.22	0.45	0.33	0.22	0.45
Dialister	0.31	0.17	0.52	0.33	0.18	0.50	0.32	0.17	0.51	0.32	0.17	0.51	0.33	0.17	0.50
Haemophilus	0.33	0.19	0.48	0.35	0.20	0.45	0.33	0.18	0.49	0.34	0.18	0.49	0.33	0.19	0.48
Escherichia	0.80	0.01	0.19	0.76	0.01	0.22	0.80	0.01	0.19	0.78	0.01	0.21	0.79	0.02	0.20
Veillonella	0.57	0.06	0.37	0.60	0.06	0.34	0.59	0.07	0.35	0.56	0.06	0.37	0.59	0.07	0.35
Coprobacillus	0.32	0.22	0.46	0.32	0.21	0.47	0.31	0.21	0.47	0.33	0.20	0.47	0.33	0.21	0.46

Table A.5: Posterior probabilities with 5 starting points. Results are from genera level IBD data with uniform prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Anaerostipes	0.32	0.24	0.43	0.33	0.23	0.44	0.32	0.25	0.43	0.34	0.23	0.43	0.32	0.25	0.43
Rothia	0.57	0.08	0.35	0.58	0.07	0.35	0.57	0.08	0.35	0.59	0.08	0.33	0.57	0.07	0.35
Klebsiella	0.57	0.07	0.36	0.56	0.07	0.37	0.55	0.07	0.38	0.55	0.07	0.38	0.56	0.07	0.37

Table A.6: Posterior probabilities with 5 starting points. Results are from genera level IBD data with sparse prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Prevotella	0.00	0.91	0.09	0.00	0.91	0.09	0.00	0.89	0.11	0.00	0.90	0.10	0.00	0.90	0.10
Faecalibacterium	0.11	0.03	0.87	0.10	0.03	0.87	0.11	0.03	0.86	0.10	0.03	0.87	0.11	0.03	0.86
Eubacterium	0.03	0.12	0.85	0.03	0.12	0.85	0.03	0.12	0.84	0.04	0.12	0.84	0.04	0.12	0.85
Alistipes	0.04	0.08	0.88	0.04	0.08	0.87	0.03	0.09	0.88	0.04	0.09	0.87	0.03	0.09	0.88
Bacteroides	0.09	0.05	0.86	0.09	0.05	0.86	0.09	0.04	0.87	0.09	0.04	0.87	0.08	0.05	0.87
Odoribacter	0.02	0.15	0.84	0.02	0.17	0.81	0.02	0.15	0.83	0.02	0.15	0.83	0.01	0.15	0.83
Dorea	0.09	0.04	0.87	0.09	0.04	0.87	0.09	0.04	0.87	0.08	0.04	0.87	0.09	0.04	0.86
Ruminococcus	0.05	0.10	0.86	0.05	0.10	0.85	0.05	0.10	0.85	0.05	0.10	0.85	0.05	0.10	0.85
Parabacteroides	0.04	0.06	0.89	0.05	0.07	0.89	0.04	0.08	0.88	0.05	0.07	0.89	0.05	0.07	0.88
Akkermansia	0.02	0.17	0.81	0.02	0.16	0.83	0.02	0.17	0.81	0.02	0.17	0.82	0.02	0.16	0.82
Roseburia	0.02	0.18	0.81	0.01	0.18	0.81	0.02	0.17	0.81	0.01	0.18	0.80	0.02	0.19	0.79
Bilophila	0.06	0.04	0.89	0.07	0.04	0.88	0.06	0.05	0.89	0.07	0.05	0.89	0.07	0.05	0.88
Streptococcus	0.11	0.03	0.86	0.12	0.03	0.85	0.12	0.03	0.85	0.12	0.03	0.84	0.11	0.03	0.86
Coprococcus	0.06	0.06	0.88	0.05	0.06	0.89	0.06	0.06	0.88	0.05	0.06	0.89	0.05	0.06	0.89
Collinsella	0.04	0.07	0.88	0.05	0.08	0.87	0.05	0.07	0.88	0.05	0.08	0.87	0.05	0.08	0.87
Clostridium	0.08	0.05	0.86	0.07	0.06	0.87	0.08	0.05	0.87	0.08	0.05	0.87	0.07	0.06	0.87
Sutterella	0.21	0.01	0.77	0.21	0.01	0.78	0.20	0.01	0.79	0.21	0.01	0.78	0.20	0.01	0.79
Bifidobacterium	0.04	0.06	0.90	0.04	0.07	0.89	0.04	0.06	0.90	0.04	0.06	0.90	0.04	0.06	0.90
Anaerotruncus	0.07	0.07	0.86	0.06	0.07	0.87	0.07	0.07	0.86	0.07	0.07	0.86	0.06	0.07	0.87
Lactococcus	0.11	0.03	0.87	0.10	0.03	0.87	0.10	0.03	0.88	0.09	0.03	0.88	0.10	0.03	0.88

Table A.6: Posterior probabilities with 5 starting points. Results are from genera level IBD data with sparse prior.

	starting 1			starting 2			starting 3			starting 4			starting 5		
set	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0	z_+	z_-	z_0
Gordonibacter	0.20	0.02	0.78	0.19	0.02	0.78	0.21	0.03	0.76	0.21	0.02	0.77	0.19	0.02	0.78
Eggerthella	0.25	0.01	0.74	0.25	0.01	0.74	0.25	0.01	0.74	0.26	0.01	0.73	0.26	0.01	0.73
Holdemania	0.09	0.04	0.87	0.08	0.04	0.88	0.08	0.05	0.88	0.09	0.04	0.87	0.08	0.04	0.87
Dialister	0.07	0.02	0.91	0.07	0.02	0.91	0.07	0.02	0.91	0.07	0.02	0.90	0.07	0.02	0.91
Haemophilus	0.11	0.02	0.87	0.12	0.02	0.86	0.11	0.02	0.87	0.11	0.02	0.86	0.11	0.02	0.87
Escherichia	0.59	0.00	0.41	0.59	0.00	0.41	0.59	0.00	0.41	0.55	0.00	0.45	0.57	0.00	0.43
Veillonella	0.24	0.01	0.75	0.23	0.01	0.77	0.22	0.01	0.77	0.23	0.01	0.76	0.24	0.01	0.75
Coprobacillus	0.11	0.03	0.86	0.11	0.03	0.86	0.11	0.03	0.86	0.11	0.03	0.86	0.11	0.03	0.86
Anaerostipes	0.13	0.04	0.83	0.12	0.04	0.84	0.12	0.04	0.84	0.12	0.04	0.84	0.12	0.04	0.84
Rothia	0.27	0.01	0.73	0.29	0.01	0.70	0.29	0.01	0.70	0.29	0.01	0.71	0.29	0.01	0.70
Klebsiella	0.31	0.01	0.68	0.28	0.01	0.71	0.31	0.01	0.68	0.34	0.01	0.65	0.32	0.01	0.68

APPENDIX B

ASSUMPTIONS AND DERIVATION OF INDIRECT EFFECT

B.1. Notation and assumptions in Chapter 4

The proposed Bayesian balance mediation model is given by

$$\boldsymbol{x_i} \sim \text{ Multinomial}(n_i, \boldsymbol{p_i}) \tag{B.1}$$

$$\boldsymbol{p_i} = \phi^{-1}(\boldsymbol{\mu_0}) \oplus \boldsymbol{a}^{t_i} \oplus \boldsymbol{U_i} \tag{B.2}$$

$$y_i|\boldsymbol{z} = a_z + c_z \cdot t_i + b_z \cdot B_{zi} + e_{zi} \tag{B.3}$$

In this model, only $\boldsymbol{x_i}, t_i, y_i$ are observed variables. As the mediator $B_{zi}$is a linear function of $\log \boldsymbol{p_i}$ for a fixed indicator vector $\boldsymbol{z}$, we use $\log \boldsymbol{p_i}$ in stating the assumptions. For a given indicator $\boldsymbol{z}$ we can replace $\log \boldsymbol{p_i}$ with B_{zi} in these assumptions and the resulting statements still hold. Using potential outcome framework, we define $\boldsymbol{p_i}(t)$ as the potential outcome for the relative abundance under $T_i = t$; $y_i(t, \log \boldsymbol{p})$ is the potential outcome under $T_i = t, \boldsymbol{p_i} = \boldsymbol{p}$. The actual latent variable $\boldsymbol{p_i} = \boldsymbol{p_i}(t_i)$ and observed outcome is expressed as $y_i = y_i(t_i, \boldsymbol{p_i})$. In this notation, T_i is the random treatment and $\boldsymbol{p_i}$ is the random relative abundance vector for subject i.

The assumptions required to derive the mediation effect (Lemma 1) are given as follows:

$$0 < P(T_i = t) < 1 \tag{B.4}$$

$$0 < P(\boldsymbol{p_i} = \boldsymbol{p}|T_i = t) < 1 \tag{B.5}$$

$$\{y_i(t', \log p), \boldsymbol{p_i}(t)\} \perp\!\!\!\perp T_i \tag{B.6}$$

$$y_i(t', \log p) \perp\!\!\!\perp \boldsymbol{p_i}(t)|T_i \tag{B.7}$$

No interaction between T_i and $\boldsymbol{p_i}$ (B.8)

No interference among subjects (B.9)

B.2. Proof of Lemma 1 in Chapter 4

Let $bal(\boldsymbol{z}, \cdot)$ be the balance index from a compositional vector and a fixed index vector $\boldsymbol{z}$. Rewriting Equations (B.2) -(B.3) in terms of $bal(\boldsymbol{z}, \cdot)$ and potential outcomes, we get

$$bal(\boldsymbol{z}, \boldsymbol{p_i}(T_i)) = bal(\boldsymbol{z}, \phi^{-1}(\boldsymbol{\mu_0})) + T_i \times bal(\boldsymbol{z}, \boldsymbol{a}) + bal(\boldsymbol{z}, \boldsymbol{U_i}(T_i)) \tag{B.10}$$

$$y_i(T_i, \log \boldsymbol{p_i}(T_i)) = a_z + c_z \times T_i + b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i}(T_i)) + e_{zi}(T_i, \log \boldsymbol{p_i}(T_i)). \tag{B.11}$$

Assumption (B.6) implies $\log \boldsymbol{U_i}(t_i) \perp\!\!\!\perp T_i$ and $bal(\boldsymbol{z}, \boldsymbol{U_i}(t_i)) \perp\!\!\!\perp T_i$ with a fixed $\boldsymbol{z}$. As a result,

$$\mathbb{E}\left(bal(\boldsymbol{z}, \boldsymbol{U_i}(T_i))|T_i = t_i\right) = \mathbb{E}\left(bal(\boldsymbol{z}, \boldsymbol{U_i}(t_i))\right) = 0 \tag{B.12}$$

which leads to

$$\mathbb{E}\left(bal(\boldsymbol{z}, \boldsymbol{p_i})|T_i\right) = bal(\boldsymbol{z}, \phi^{-1}(\boldsymbol{\mu_0})) + T_i \times bal(\boldsymbol{z}, \boldsymbol{a}) \tag{B.13}$$

Assumption (B.7) implies $e_{zi}(t_i, \log \boldsymbol{p_i}) \perp\!\!\!\perp bal_z\left(\boldsymbol{p_i}(T_i)\right)|T_i = t_i$ and leads to

$$\mathbb{E}\left(e_{zi}(T_i, \log \boldsymbol{p_i}(T_i)|T_i = t_i, \log \boldsymbol{p_i}(T_i) = \log \boldsymbol{p_i}\right) = \mathbb{E}\left(e_{zi}(T_i, \log \boldsymbol{p_i})|T_i = t_i\right). \tag{B.14}$$

Under the assumption (B.6), the equation can be further simplified as $\mathbb{E}\left(e_{zi}(t_i, \log \boldsymbol{p_i})\right) = 0$ and applying this to Equation (B.11), we have

$$\mathbb{E}\left(y_i|\log \boldsymbol{p_i}, T_i\right) = a_z + c_z \times T_i + b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i}) = \mathbb{E}\left(y_i|bal(\boldsymbol{z}, \boldsymbol{p_i}), T_i\right) \tag{B.15}$$

Based on these assumptions, the causal direct effect can be derived as

$$
\begin{aligned}
\zeta(\tau) =& \mathbb{E}y_i(t, \log \boldsymbol{p_i(\tau)}) - \mathbb{E}y_i(t_0, \log \boldsymbol{p_i(\tau)}) \\
&= \int [\mathbb{E}(y_i(t, \log \boldsymbol{p})| \log \boldsymbol{p_i(\tau)} = \log \boldsymbol{p}) - \\
&\quad \mathbb{E}(y_i(t_0, \log \boldsymbol{p})| \log \boldsymbol{p_i(\tau)} = \log \boldsymbol{p})] \, dF_{\log \boldsymbol{p_i(\tau)}}(\log \boldsymbol{p}) \\
&= \int [\mathbb{E}(y_i(t, \log \boldsymbol{p})| \log \boldsymbol{p_i(\tau)} = \log \boldsymbol{p}, T_i = \tau) - \\
&\quad \mathbb{E}(y_i(t_0, \log \boldsymbol{p})| \log \boldsymbol{p_i(\tau)} = \log \boldsymbol{p}, T_i = \tau)] \, dF_{\log \boldsymbol{p_i(\tau)}}(\log \boldsymbol{p}) \\
&= \int [\mathbb{E}(y_i(t, \log \boldsymbol{p})|T_i = \tau) - \ \mathbb{E}(y_i(t_0, \log \boldsymbol{p})|T_i = \tau)] \, dF_{\log \boldsymbol{p_i(\tau)}}(\log \boldsymbol{p}) \\
&= \int [\mathbb{E}(y_i(t, \log \boldsymbol{p})|T_i = t) - \ \mathbb{E}(y_i(t_0, \log \boldsymbol{p})|T_i = t_0)] \, dF_{\log \boldsymbol{p_i}|T_i=\tau}(\log \boldsymbol{p}) \\
&= \int [\mathbb{E}(y_i(t, \log \boldsymbol{p})| \log \boldsymbol{p_i(t)} = \log \boldsymbol{p}, T_i = t) - \\
&\quad \mathbb{E}(y_i(t_0, \log \boldsymbol{p})|, \log \boldsymbol{p_i(t_0)} = \log \boldsymbol{p}, T_i = t_0)] \, dF_{\log \boldsymbol{p_i}|T_i=\tau}(\log \boldsymbol{p}) \\
&= \int [\mathbb{E}(y_i| \log \boldsymbol{p_i(t)} = \log \boldsymbol{p}, T_i = t) - \\
&\quad \mathbb{E}(y_i| \log \boldsymbol{p_i(t_0)} = \log \boldsymbol{p}, T_i = t_0)] \, dF_{\log \boldsymbol{p_i}|T_i=\tau}(\log \boldsymbol{p}) \\
&\quad = \int c_z \times (t - t0) dF_{\log \boldsymbol{p_i}|T_i=\tau}(\log \boldsymbol{p}) \\
&\quad = c_z \times (t - t0)
\end{aligned}
$$

where F_{v1} and $F_{v1|v2}$ are the cumulative distribution functions of $v1$ and $v1$ conditioning on the value of $v2$

We can prove that the indirect effect, i.e., the mediation effect in a similar fashion.

$$
\begin{aligned}
\delta(\tau) =& \mathbb{E}y_i(\tau, \log \boldsymbol{p_i(t)}) - \mathbb{E}y_i(\tau, \log \boldsymbol{p_i(t_0)}) \\
=& \int \mathbb{E}(y_i(\tau, \log \boldsymbol{p})|T_i = \tau, \log \boldsymbol{p_i}(t) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t)}}(\log \boldsymbol{p}) - \\
& \mathbb{E}(y_i(\tau, \log \boldsymbol{p})|T_i = \tau, \log \boldsymbol{p_i}(t_0) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t_0)}}(\log \boldsymbol{p}) \\
=& \int \mathbb{E}(y_i(t, \log \boldsymbol{p})|T_i = \tau, \log \boldsymbol{p_i}(t) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t)}}(\log \boldsymbol{p}) - \\
& \mathbb{E}(y_i(t_0, \log \boldsymbol{p})|T_i = \tau, \log \boldsymbol{p_i}(t_0) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t_0)}}(\log \boldsymbol{p}) \\
=& \int \mathbb{E}(y_i(t, \log \boldsymbol{p})|T_i = t, \log \boldsymbol{p_i}(t) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t)}}(\log \boldsymbol{p}) - \\
& \mathbb{E}(y_i(t_0, \log \boldsymbol{p})|T_i = t_0, \log \boldsymbol{p_i}(t_0) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t_0)}}(\log \boldsymbol{p}) \\
=& \int \mathbb{E}(y_i|T_i = t, \log \boldsymbol{p_i}(t) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t)}}(\log \boldsymbol{p}) - \\
& \mathbb{E}(y_i|T_i = t_0, \log \boldsymbol{p_i}(t_0) = \log \boldsymbol{p}) dF_{\log \boldsymbol{p_i(t_0)}}(\log \boldsymbol{p}) \\
=& \int (a_z + c_z \times t + b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i(t)})) dF_{\log \boldsymbol{p_i(t)}}(\log \boldsymbol{p}) - \\
& (a_z + c_z \times t_0 + b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i(t_0)})) dF_{\log \boldsymbol{p_i(t_0)}}(\log \boldsymbol{p}) \\
=& \int b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i(t)}) dF_{bal(\boldsymbol{z}, \boldsymbol{p_i(t)})} - \int b_z \times bal(\boldsymbol{z}, \boldsymbol{p_i(t_0)}) dF_{bal(\boldsymbol{z}, \boldsymbol{p_i(t_0)})} \\
=& b_z \times [\mathbb{E}\left(bal(\boldsymbol{z}, \boldsymbol{p_i})|T_i = t\right) - \mathbb{E}\left(bal(\boldsymbol{z}, \boldsymbol{p_i})|T_i = t_0\right)] \\
=& b_z \times (t - t_0) \times bal(\boldsymbol{z}, a)
\end{aligned}
$$

In two group analysis where T_i only has two values $0, 1$, the perturbation vector is expressed as $\phi^{-1}(\mu_1 - \mu_0)$. The mediation effect is simplified as $b_z \times bal(\boldsymbol{z}, \phi^{-1}(\mu_1 - \mu_0))$.

BIBLIOGRAPHY

Aitchison, J (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, 139–177.

Aitchson, J (2003). *The Statistical Analysis of Compositional Data*. The Blackburn Press.

Albert, JH and Chib, S (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* 88.422, 669–679.

Backhed, F, Ding, H, Wang, T, Hooper, LV, Koh, GY, Nagy, A, Semenkovich, CF, and Gordon, JI (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America* 101.44, 15718.

Bates, S and Tibshirani, R (2019). Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biometrics* 75.2, 613–624.

Besten, G den, Eunen, K van, Groen, AK, Venema, K, Reijngoud, D-J, and Bakker, BM (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of lipid research* 54.9, 2325–2340.

Billheimer, D, Guttorp, P, and Fagan, W (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association* 96, 1205–1214.

Bollen, KA and Stine, R (1990). Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability. *Sociological Methodology* 20, 115–140.

Bryrup, T, Thomsen, C, Kern, T, Allin, K, Brandslund, I, Jrgensen, N, Vestergaard, H, Hansen, T, Hansen, T, Pedersen, O, and Nielsen, T (2019). Metformin-induced changes of the gut microbiota in healthy young men: results of a non-blinded, one-armed intervention study. *Diabetologia* 62, 1024–1035.

Cao, Y, Lin, W, and Li, H (2017). Two-sample tests of high-dimensional means for compositional data. *Biometrika* 105.1, 115–132.

Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, Fierer, N, Pena, AG, Goodrich, JK, Gordon, JI, Huttley, GA, Kelley, ST, Knights, D, Koenig, JE, Ley, RE, Lozupone, CA, McDonald, D, Muegge, BD, Pirrung, M, Reeder, J, Sevinsky, JR, Turnbaugh, PJ, Walters, WA, Widmann, J, Yatsunenko, T, Zaneveld, J, and Knight, R (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 7.5, 335–336.

Castaner, O, Goday, A, Park, Y-M, Lee, S-H, Magkos, F, Shiow, S-ATE, and Schröder, HE (2018). The Gut Microbiome Profile in Obesity: A Systematic Review. *International Journal of Endocrinology* 2018, 4095789 KW –.

Chén, OY, Crainiceanu, C, Ogburn, EL, Caffo, BS, Wager, TD, and Lindquist, MA (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* 19.2, 121–136.

Cheung, MWL (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods* 41.2, 425–438.

Cryan, JF, O'Riordan, KJ, Cowan, CSM, Sandhu, KV, Bastiaanssen, TFS, Boehme, M, Codagnone, MG, Cussotto, S, Fulling, C, Golubeva, AV, Guzzetta, KE, Jaggar, M, Long-Smith, CM, Lyte, JM, Martin, JA, Molinero-Perez, A, Moloney, G, Morelli, E, Morillas, E, O'Connor, R, Cruz-Pereira, JS, Peterson, VL, Rea, K, Ritz, NL, Sherwin, E, Spichak, S, Teichman, EM, Wouw, M van de, Ventura-Silva, AP, Wallace-Fitzsimons, SE, Hyland, N, Clarke, G, and Dinan, TG (2019). The Microbiota-Gut-Brain Axis. *Physiological Reviews* 99.4, 1877–2013.

Das, B and Nair, GB (2019). Homeostasis and dysbiosis of the gut microbiome in health and disease. *Journal of Biosciences* 44.5, 117.

DeSantis, TZ, Hugenholtz, P, Larsen, N, Rojas, M, Brodie, EL, Keller, K, Huber, T, Dalevi, D, Hu, P, and Andersen, GL (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* 72.7, 5069–5072.

Egozcue, JJ and Pawlowsky-Glahn, V (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology* 37.7, 795–828.

Gelfand, AE, Hills, SE, Racine-Poon, A, and Smith, AFM (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association* 85.412, 972–985.

George, EI and McCulloch, RE (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* 88.423, 881–889.

Goodrich, JK, Davenport, ER, Beaumont, M, Jackson, MA, Knight, R, Ober, C, Spector, TD, Bell, JT, Clark, AG, and Ley, RE (2016). Genetic Determinants of the Gut Microbiome in UK Twins. *Cell host & microbe* 19, 731–743.

Griffin, AS, West, SA, and Buckling, A (2004). Cooperation and competition in pathogenic bacteria. *Nature* 430.7003, 1024–1027.

Halfvarson, J, Brislawn, CJ, Lamendella, R, Vázquez-Baeza, Y, Walters, WA, Bramer, LM, D'Amato, M, Bonfiglio, F, McDonald, D, Gonzalez, A, McClure, EE, Dunklebarger, MF, Knight, R, and Jansson, JK (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology* 2, 17004–17004.

Herbrand, H, Bernhardt, G, Forster, R, and Pabst, O (2008). Dynamics and Function of Solitary Intestinal Lymphoid Tissue. *Critical Reviews in Immunology* 28.1, 1–13. ISSN: 1040-8401.

Holmes, CC and Held, L (2006). Bayesian auxiliary variable models for binary and multinomial regression. 1.1, 145–168.

Holmes, E, Li, JV, Athanasiou, T, Ashrafian, H, and Nicholson, JK (2011). Understanding the role of gut microbiome host metabolic signal disruption in health and disease. *Trends in Microbiology* 19.7, 349–359.

Hooper, LV (2009). Do symbiotic bacteria subvert host immunity? *Nature Reviews Microbiology* 7.5, 367–374.

Hu, FB, Rimm, E, Smith-Warner, SA, Feskanich, D, Stampfer, MJ, Ascherio, A, Sampson, L, and Willett, WC (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *The American journal of clinical nutrition* 69.2, 243–249.

Huang, L and Li, H (2020). Bayesian balance regression in microbiome studies using stochastic search. *submitted*.

Imai, K, Keele, L, and Yamamoto, T (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. 25.1, 51–71.

Kassam, Z, Lee, CH, Yuan, Y, and Hunt, RH (2013). Fecal Microbiota Transplantation forClostridium difficileInfection: Systematic Review and Meta-Analysis. *American Journal of Gastroenterology KW* - 108.4.

Kim, O-S, Cho, Y-J, Lee, K, Yoon, S-H, Kim, M, Na, H, Park, S-C, Jeon, YS, Lee, J-H, Yi, H, Won, S, and Chun, J (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International journal of systematic and evolutionary microbiology* 62.Pt 3, 716–721.

Kurtz, ZD, Müller, CL, Miraldi, ER, Littman, DR, Blaser, MJ, and Bonneau, RA (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS computational biology* 11.5, e1004226.

Lewis, JD, Chen, EZ, Baldassano, RN, Otley, AR, Griffiths, AM, Lee, D, Bittinger, K, Bailey, A, Friedman, ES, Hoffmann, C, Albenberg, L, Sinha, R, Compher, C, Gilroy, E, Nessel, L, Grant, A, Chehoud, C, Li, H, Wu, GD, and Bushman, FD (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell host & microbe* 18.4, 489–500.

Ley, RE, Peterson, DA, and Gordon, JI (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell* 124.4, 837–848.

Ley, RE, Backhed, F, Turnbaugh, P, Lozupone, CA, Knight, RD, and Gordon, JI (2005a). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* 102.31, 11070–11075.

Ley, RE, Backhed, F, Turnbaugh, P, Lozupone, CA, Knight, RD, and Gordon, JI (2005b). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* 102.31, 11070.

Ley, RE, Turnbaugh, PJ, Klein, S, and Gordon, JI (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444.7122, 1022–1023.

Li, H (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annu. Rev. Stat. Appl.* 2.1, 73–94.

Lu, J, Shi, P, and Li, H (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* 75.1, 235–244.

Morgan, JL, Darling, AE, and Eisen, JA (2010). Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *PLOS ONE* 5.4, e10209.

Morrison, DJ and Preston, T (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut microbes* 7.3, 189–200.

Morton, JT, Sanders, J, Quinn, RA, McDonald, D, Gonzalez, A, Vázquez-Baeza, Y, Navas-Molina, JA, Song, SJ, Metcalf, JL, Hyde, ER, Lladser, M, Dorrestein, PC, and Knight, R (2017). Balance Trees Reveal Microbial Niche Differentiation. *mSystems* 2.1, e00162–16.

Pearl, J (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Rivera-Pinto, J, Egozcue, JJ, Pawlowsky-Glahn, V, Paredes, R, Noguera-Julian, M, and Calle, ML (2018). Balances: a New Perspective for Microbiome Analysis. *mSystems* 3.4, e00053–18.

Rubin, DB (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association* 100.469, 322–331.

Savage, DC (1977). MICROBIAL ECOLOGY OF THE GASTROINTESTINAL TRACT. *Annual Review of Microbiology* 31.1. PMID: 334036, 107–133. DOI: 10.1146/annurev.mi.31.100177.000543. eprint: https://doi.org/10.1146/annurev.mi.31.100177.000543. URL: https://doi.org/10.1146/annurev.mi.31.100177.000543.

Segata, N, Waldron, L, Ballarini, A, Narasimhan, V, Jousson, O, and Huttenhower, C (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9.8, 811–814.

Sekirov, I, Russell, SL, Antunes, LCM, and Finlay, BB (2010). Gut Microbiota in Health and Disease. *Physiological Reviews* 90.3, 859–904.

Shi, P, Zhang, A, and Li, H (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* 10.2, 1019–1040.

Shrout, PE and Bolger, N (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological methods* 7.4, 422–445.

Sobel, ME (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology* 13, 290–312.

Sohn, MB and Li, H (2019). Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics* 13.1, 661–681.

Tomova, A, Bukovsky, I, Rembert, E, Yonas, W, Alwarith, J, Barnard, ND, and Kahleova, H (2019). The Effects of Vegetarian and Vegan Diets on Gut Microbiota. *Frontiers in Nutrition* 6, 47.

Turnbaugh, PJ, Hamady, M, Yatsunenko, T, Cantarel, BL, Duncan, A, Ley, RE, Sogin, ML, Jones, WJ, Roe, BA, Affourtit, JP, Egholm, M, Henrissat, B, Heath, AC, Knight, R, and Gordon, JI (2009). A core gut microbiome in obese and lean twins. *Nature* 457.7228, 480–484.

VanderWeele, TJ and Vansteelandt, S (2014). Mediation Analysis with Multiple Mediators. *Epidemiologic methods* 2.1, 95–115.

Vétizou, M, Pitt, JM, Daillère, R, Lepage, P, Waldschmitt, N, Flament, C, Rusakiewicz, S, Routy, B, Roberti, MP, Duong, CPM, Poirier-Colame, V, Roux, A, Becharef, S, Formenti, S, Golden, E, Cording, S, Eberl, G, Schlitzer, A, Ginhoux, F, Mani, S, Yamazaki, T, Jacquelot, N, Enot, DP, Bérard, M, Nigou, J, Opolon, P, Eggermont, A, Woerther, P-L, Chachaty, E, Chaput, N, Robert, C, Mateus, C, Kroemer, G, Raoult, D, Boneca, IG, Carbonnel, F, Chamaillard, M, and Zitvogel, L (2015). Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science (New York, N.Y.)* 350.6264, 1079–1084.

Wang, T and Zhao, H (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics* 11.2, 771–791.

Washburne, AD, Silverman, JD, Leff, JW, Bennett, DJ, Darcy, JL, Mukherjee, S, Fierer, N, and David, LA (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* 5, e2969–e2969.

West, SA, Griffin, AS, Gardner, A, and Diggle, SP (2006). Social evolution theory for microorganisms. *Nature Reviews Microbiology* 4.8, 597–607.

Whitman, WB, Coleman, DC, and Wiebe, WJ (1998). Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* 95.12, 6578.

Wu, GD, Chen, J, Hoffmann, C, Bittinger, K, Chen, Y-Y, Keilbaugh, SA, Bewtra, M, Knights, D, Walters, WA, Knight, R, Sinha, R, Gilroy, E, Gupta, K, Baldassano, R, Nessel, L, Li, H, Bushman, FD, and Lewis, JD (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)* 334.6052, 105–108.

Wu, GD, Compher, C, Chen, EZ, Smith, SA, Shah, RD, Bittinger, K, Chehoud, C, Albenberg, LG, Nessel, L, Gilroy, E, Star, J, Weljie, AM, Flint, HJ, Metz, DC, Bennett, MJ, Li, H, Bushman, FD, and Lewis, JD (2016). Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut* 65.1, 63–72. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2014-308209. eprint: https://gut.bmj.com/content/65/1/63.full.pdf. URL: https://gut.bmj.com/content/65/1/63.

Xu, J, Nelson, B, and Hong, L (2013). An Adaptive Hyperbox Algorithm for High-Dimensional Discrete Optimization via Simulation Problems. English. *INFORMS Journal on Computing* 25.1, 133–146. ISSN: 1091-9856. DOI: 10.1287/ijoc.1110.0481.

Zhang, H, DiBaise, JK, Zuccolo, A, Kudrna, D, Braidotti, M, Yu, Y, Parameswaran, P, Crowell, MD, Wing, R, Rittmann, BE, and Krajmalnik-Brown, R (2009). Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences of the United States of America* 106.7, 2365–2370.