

STATISTICAL METHODS FOR MODELING COMPLEX DEPENDENCY STRUCTURES IN  
ZERO-INFLATED METAGENOMIC SEQUENCING DATA

Rebecca Ann Deek

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

2023

Supervisor of Dissertation

Hongzhe Li, Perelman Professor of Biostatistics, Epidemiology and Informatics

Graduate Group Chairperson

Russell T. Shinohara, Professor of Biostatistics

Dissertation Committee

Mingyao Li, Professor of Biostatistics

Jing Huang, Assistant Professor of Biostatistics

Ronald G. Collman, Professor of Medicine

STATISTICAL METHODS FOR MODELING COMPLEX DEPENDENCY STRUCTURES  
IN ZERO-INFLATED METAGENOMIC SEQUENCING DATA

COPYRIGHT

2023

Rebecca Ann Deek

This work is licensed under the  
Creative Commons Attribution  
NonCommercial-ShareAlike 4.0  
License

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

## ACKNOWLEDGEMENT

The mentorship, support, and generosity of many have enabled the materialization of this dissertation. I am filled with deep gratitude for all of you.

First and foremost, I would like to thank my advisor, Dr. Hongzhe Li. Your guidance, mentorship, and unwavering support over the last five years were pivotal in carrying out this research. From these, I have learned more about scholarship, presentation, critical analysis, and becoming an independent researcher. You have been an exceptional teacher both in the classroom and in our many meetings. Your enthusiasm for our work always left me with a better understanding and renewed energy to solve the problem at hand. It has been a privilege to work with you.

I would like to express my sincere appreciation to my committee members: Drs. Mingyao Li, Jing Huang, and Ron Collman, who generously dedicated time toward guiding me through this work. Each of you brought a different perspective and expertise to our meetings. Your insightful questions and thoughtful feedback prompted me to think more critically about the problems I was working on, as well as become a better researcher. I would also like to thank Dr. Zhi Wei for his mentorship and advice, which propelled me down this path, since I was an undergraduate student. This work has been financially supported by the National Institutes of Health through the GM123056 and GM129781 grants.

It has been an absolute pleasure getting to know the current and former GGEB students. Thank you for the coffee chats, lunch breaks, and long conversations. These were among the highlights of my graduate school experience at Penn. Without all of you, my time in this program would not have been nearly as memorable. In particular, my cohort— Francesca, Sarah, Danni, Andrew, Melissa, Haotian, and Elle— has been there to celebrate, commiserate, and motivate each other from the beginning. I am grateful for your friendship and to have been surrounded by such smart, kind, and thoughtful people for the last five years.

To my friends outside of this program, thank you for being a part of my support system. You have

been a constant source of joy and always reminded me of the importance of taking breaks. You all have brightened my life.

Finally, to my family, without whom I am certain I would not have made it to this point. To my parents, Maura and Fadi, thank you for your unconditional love, support, and faith in me. This accomplishment would not have been possible without you and your sacrifices. To my brothers, Matthew and Andrew, thank you for your love, encouragement, and advice. Your own drive and passion have motivated me for as long as I can remember. To my sister-in-law, Michelle, thank you for your encouragement and understanding through the highs and lows. Finally, I thank my grandparents, who laid the foundation for this work by instilling the importance of education into each generation.

## ABSTRACT

### STATISTICAL METHODS FOR MODELING COMPLEX DEPENDENCY STRUCTURES IN ZERO-INFLATED METAGENOMIC SEQUENCING DATA

Rebecca Ann Deek

Hongzhe Li

Advances in high-throughput sequencing technologies have enabled large-scale metagenomic sequencing studies of microbial compositions. As such, there is a growing scientific interest in understanding the human microbiome, defined as all the microorganisms and their genes in, or on, the body. Of particular interest is its functional role in human-host health. Nevertheless, there remains a statistical and computational bottleneck in effectively analyzing data from 16S rRNA and metagenomic sequencing studies. This is due to the characteristic excessive zeros, sequencing depth constraints, and high dimensionality of such data. Motivated by numerous microbiome studies, this dissertation aims to narrow the gap by developing novel statistical methods specifically designed to capture the excessive zeros of the data. The specific aims are to develop statistical models, inference procedures, and computational fast algorithms to (1) identify distinct microbial communities in a given data set, as well as each community's important bacterial taxa, and (2) build microbial covariation networks based upon the estimated covariation between a pair of zero-inflated variables. To this end, three methodological advances are proposed. First, a generative latent mixture model of microbial counts that distinguishes between structural and sampling zeros. Second, a mixture margin copula model and two-stage inference procedure for microbial covariation networks in cross-sectional studies. Third, an extension to random-effects mixture margin copula models, as well as a corresponding Monte Carlo EM algorithm and likelihood ratio test to build temporally conserved covariation networks from longitudinal data. Furthermore, the performance and utility of these methods are demonstrated using simulations and several publicly available microbiome data sets.

# TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iii
ABSTRACT . . . . .	v
LIST OF TABLES . . . . .	viii
LIST OF ILLUSTRATIONS . . . . .	ix
CHAPTER 1 : INTRODUCTION . . . . .	1
CHAPTER 2 : A ZERO-INFLATED LATENT DIRICHLET ALLOCATION MODEL FOR MICROBIOME STUDIES . . . . .	5
2.1 Introduction . . . . .	5
2.2 Latent Variable Mixture Modeling for Microbiome Studies . . . . .	7
2.3 A Collapsed Gibbs Sampler for Model Inference . . . . .	10
2.4 Simulation Studies . . . . .	13
2.5 Identification of Gut Microbial Subcommunities in a Healthy Cohort . . . . .	17
2.6 Discussion . . . . .	19
CHAPTER 3 : INFERENCE OF MICROBIAL COVARIATION NETWORKS USING COPULA MODELS WITH MIXTURE MARGINS . . . . .	23
3.1 Introduction . . . . .	23
3.2 Copula Models with Mixture Margin Distributions . . . . .	25
3.3 A Two-Stage Estimation Method and Statistical Inference . . . . .	28
3.4 Simulation Studies . . . . .	32
3.5 Analysis of a Microbial Covariation Network in the Healthy Human Gut . . . . .	40
3.6 Discussion . . . . .	44

CHAPTER 4 : MIXTURE MARGIN RANDOM-EFFECTS COPULA MODELS FOR INFERRING TEMPORALLY CONSERVED MICROBIAL COVARIATION NETWORKS FROM LONGITUDINAL DATA . . . . .	46
4.1 Introduction . . . . .	46
4.2 Mixture margin random-effects copula models and Monte Carlo inference . . . . .	49
4.3 Simulation studies . . . . .	57
4.4 Analysis of temporally conserved microbial networks during childhood development .	60
4.5 Discussion . . . . .	67
CHAPTER 5 : DISCUSSION . . . . .	70
APPENDIX A : SUPPLEMENTARY MATERIALS FOR CHAPTER 2 . . . . .	73
APPENDIX B : SUPPLEMENTARY MATERIALS FOR CHAPTER 3 . . . . .	84
APPENDIX C : SUPPLEMENTARY MATERIALS FOR CHAPTER 4 . . . . .	94
BIBLIOGRAPHY . . . . .	96

## LIST OF TABLES

TABLE 2.1	Comparison of estimated structural zero taxa from the zinLDA model to true structural zero taxa from simulation across different parameter settings using sensitivity, specificity, positive predictive value, and negative predictive value. A “positive” results is assumed to be $\hat{\beta}_{ij} = 0$ . . . . .	18
-----------	--	----

## LIST OF ILLUSTRATIONS

FIGURE 2.1	Plate diagrams of (a) Latent Dirichlet Allocation and (b) zero-inflated Latent Dirichlet Allocation. Nodes represent parameters and random variables, shading denotes the observed data, and boxes represent repeated sampling. The outer box is denoted with D for once per biological sample, the inner box with N for sampling per sequencing read, and the upper box with K for per subcommunity. . . . .	9
FIGURE 2.2	Bar graphs of the top eight taxa for each of the five subcommunities alongside their corresponding $\hat{\beta}_{ij}$ values. The first column contains the ground truth (i.e., top taxa from simulation). The second and third columns are the estimated top taxa from the zinLDA and LDA models, respectively. . . . .	14
FIGURE 2.3	Observed and posterior predictive simulated <i>asinh</i> -transformed taxon counts plotted in order of increasing true abundance. Each panel is a different biological sample. The solid black line represents the true counts. The pink and blue points are the counts from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median counts across the 50 data sets. . . . .	16
FIGURE 2.4	Bar graphs of the representative taxa for each of the five subcommunities alongside their corresponding $\hat{\beta}_{ij}$ values for two independent subsets of 1000 randomly selected subjects from the American Gut Project. . . . .	20
FIGURE 3.1	Boxplots of two-stage and plug-in estimated $\tilde{\theta}$ across 500 simulations, where plug-in estimators use the true value of marginal parameters and perform univariate estimation of $\theta$ . The black dashed line represents the true $\theta$ value. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ). . . . .	34
FIGURE 3.2	Boxplots of estimated Spearman's correlation, using copula and sample estimators, across 500 simulations. The black dashed line represents the true value. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ). . . . .	36
FIGURE 3.3	Power curves of the two-stage likelihood ratio test for independence compared to sample correlation tests. Black dashed lines at 0.05 (Type I error) and 0.8. Power was calculated under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ) with covariate adjustment. . . . .	38
FIGURE 3.4	ROC curves based the two-stage estimator of the copula dependence parameter and SparCC under multivariate Gaussian copula simulation. Cutoffs were selected based on two-stage likelihood ratio test p-values for the copula model and the absolute value of the estimated correlation for SparCC, as suggested in the original manuscript. . . . .	39
FIGURE 3.5	Heatmap of the AGP adjacency matrix with three clusters identified by complete agglomerative hierarchical clustering. Red indicates positive covariation and blue negative covariation. . . . .	43

FIGURE 4.1	Scatterplots of the proportion relative abundance of two pairs of bacteria (top: <i>Veillonella</i> and <i>Bacteroides</i> , bottom: <i>Veillonella</i> and <i>Clostridium</i> ) over the first seven months of life for children unexposed to antibiotics in the DIABIMMUNE study (Yassour et al., 2016). The general covariation structure is consistent over the seven months. A negative covariation for <i>Veillonella</i> and <i>Bacteroides</i> and positive covariation for <i>Veillonella</i> and <i>Clostridium</i> , illustrating a conserved dependence structure between the taxa. . . . .	47
FIGURE 4.2	Boxplots of the $\tilde{\theta}$ and $\tilde{\sigma}$ estimates under seven different parameter ( $\theta$ and $\sigma$ ) settings in simulation and $n = 100$ . The first four are under the alternative hypothesis ( $\theta \neq 0$ for the Frank copula) and the last three are under the null of no conserved covariation structure ( $\theta = 0$ ). The black dashed line represents the mean across all runs. The red star indicates the true value specified under simulation. . . . .	59
FIGURE 4.3	Plots of the posterior mean of $\theta^{(t)}$ for a randomly selected subset of ten significant pairs, split by group (antibiotics vs. no antibiotics). Many of the pairs show relatively constant dependence parameters over time. Magnitude of the estimated posterior means and variability around the conserved value differ between groups for some pairs. . . . .	62
FIGURE 4.4	Conserved microbial covariation network diagrams for the antibiotics and no antibiotics groups. Networks were built using the estimated conserved mean dependence parameter ( $\tilde{\theta}$ ). Each node represents a bacterial genus, node shape and color correspond to graph cluster from an algorithm that maximized modularity. Two nodes share an edge if their FDR corrected p-value from the Monte Carlo likelihood ratio test was less than 0.05. Edge color corresponds to positive/negative (orange/green) dependence. The no antibiotics network has a higher edge density and lower modularity. . . . .	64
FIGURE 4.5	Node-level centrality statistics, degree and betweenness, plotted for each microbe to compare the antibiotics and no antibiotics networks. Dot size is proportional to the absolute difference in mean abundance between the two groups and node color indicates which group has larger mean abundance of the bacteria (i.e., blue indicates larger mean abundance in the no antibiotics group). Red line corresponds to $y = x$ . . . . .	66
FIGURE 4.6	Network fragility measured by diameter and efficiency as nodes are sequentially removed in targeted and random attacks. The no antibiotics network is more robust to targeted attacks which remove nodes in decreasing degree order and maintains a higher efficiency than the antibiotics network in random attacks. . . . .	68
FIGURE A.1	Bar graphs of the top eight taxa for each subcommunity with their corresponding $\hat{\beta}_{ij}$ values under model misspecification. Data was simulated under a true zero-inflated latent Dirichlet allocation model with five communities and observed $V = 87$ . An underspecified model with four (left) and an overspecified model with six (right) subcommunities were fit to the data. . . . .	80

FIGURE A.2	Observed and posterior predictive simulated <i>asinh</i> -transformed taxon counts plotted in order of increasing observed abundance from the American Gut Project. Each panel is a different biological sample. The solid black line represents the observed counts. The pink and blue points are the counts from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median counts across the 50 data sets. . . . .	81
FIGURE A.3	Observed and posterior predictive simulated relative abundances of taxa from the same six samples selected in Figure A.2, plotted in order of increasing observed relative abundance from the American Gut Project. The solid black line represents the observed relative abundance. The pink and blue points are the relative relative abundances from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median abundances across the 50 data sets. The first and third column are zoom in on columns two and four, respectively, showing relative abundances between 0 and 0.1 only. . . . .	82
FIGURE A.4	Boxplot of the posterior estimates of $\pi_{ij}$ for taxa with zero counts, split by subcommunity, from the zinLDA model applied to a subset of 1000 subjects from the American Gut Project. . . . .	83
FIGURE B.1	Boxplots of estimated $\tilde{\theta}$ values across 500 simulations. The black dashed line represents the true $\theta$ value. Data was simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ). . . . .	89
FIGURE B.2	Boxplots of the jackknife variance of two-stage and plug-in estimated $\tilde{\theta}$ , denoted as $\hat{\sigma}_{\tilde{\theta}}^2$ , across 500 simulations. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ). Outliers with variance values greater than 25 were removed from plots for visualization. . . . .	90
FIGURE B.3	Boxplots of the estimated jackknife variance of $\tilde{\theta}$ , denoted as $\hat{\sigma}_{\tilde{\theta}}^2$ , across 500 simulations. Data was simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ). Outliers with variance values greater than 50 were removed from plots for visualization. . . . .	91
FIGURE B.4	Mean and standard error bars of the estimated jackknife variance of $\tilde{\theta}$ from data simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ). Black triangles correspond the empirical (sample) variance. . . . .	92
FIGURE B.5	Boxplot of estimated Spearman's correlation, using copula and sample estimators, across 500 simulations. The black dashed line represents the true value. Data was simulated from a Gaussian copula function with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ). . . . .	93

FIGURE C.1 Boxplots of the  $\tilde{\theta}$  and  $\tilde{\sigma}$  estimates under seven different parameter ( $\theta$  and  $\sigma$ ) settings in simulation and  $n = 50$ . The first four are under the alternative hypothesis ( $\theta \neq 0$  for the Frank copula) and the last three are under the null of no conserved covariation structure ( $\theta = 0$ ). The black dashed line represents the mean across all runs. The red star indicates the true value specified under simulation. . . . . 94

FIGURE C.2 Violin plots of estimated  $\tilde{\theta}$  and  $\tilde{\sigma}$  from the random-effect for all pairs in the DIABIMMUNE data. Data is split by antibiotics exposure status and significance of the Monte Carlo likelihood ratio test. The distribution of the estimated mean ( $\tilde{\theta}$ ) is shift towards the null value of zero for non-significant pairs. The distribution of the estimated standard deviation ( $\tilde{\sigma}$ ) is similar across significance levels. These trends are consistent across antibiotics exposure status. 95

# CHAPTER 1

## INTRODUCTION

The microbiome, which refers to all the microorganisms and their genes in a well-defined environment, has long been of interest in biomedical research (Burge, 1988; Lederberg and Mccray, 2001). The advent and proliferation of Next-Generation Sequencing (NGS) technologies has given rise to many large-scale high-throughput microbiome studies. These include The Human Microbiome Project (Turnbaugh et al., 2007) and American Gut Project (McDonald et al., 2018), both of which focused on characterizing the microbiome of predominantly healthy subjects. As well as the Earth Microbiome Project that aimed to describe the uncultured diversity of the planet (Gilbert et al., 2014). Data from these studies provide important information on taxonomic classification and microbial diversity. Differential abundance analysis and microbiome association tests have been widely used to understand how the host environment (e.g. human-host health) is associated with the microbiome (Paulson et al., 2013; Peng et al., 2015; Scealy and Welsh, 2011; White et al., 2009). From such work, it is now known that the human microbiome is associated with many conditions, including obesity, inflammatory bowel disease, and rheumatoid arthritis (Greenblum et al., 2012; Scher and Abramson, 2011; Taneja, 2014). Similarly, salinity, ecosystem type, and pH are important factors in determining soil microbial composition (Lozupone and Knight, 2007; Fierer and Jackson, 2006; Thompson et al., 2017). This dissertation develops statistical and machine learning models, inference procedures, and efficient computational algorithms for microbial community and covariation analysis.

Statistical and computational analysis of 16S rRNA and shotgun metagenomic sequencing data requires careful thought and consideration for several reasons. The data from such studies are often noisy, heterogeneous, and reflect relative, not absolute, abundances of the bacterial taxa in the environment. One of the most defining and difficult features of the data are its zeros, in some data sets accounting for over 90% of the observed counts. Many existing methods aim to “Gaussian-ize” the data by performing log-ratio transformations, and then apply standard statistical

techniques. Such transformations require the use of pseudocounts, a small non-zero count added to each observation to avoid any problems with taking a log-based transformation. The problem with pseudocounts are two-fold. First, downstream analyses can be sensitive to them and give different results based on the arbitrary choice. Second, because of the excessive zeros, adding a pseudocount doesn't solve the spike in the data at zero, but instead moves it to a new value in the transformed data. Therefore, the transformed data is not Gaussian or even sub-Gaussian, limiting the use of such transformations in practice.

It is necessary for statistical techniques to explicitly account for the zeros of the data rather than treating them as an afterthought. This dissertation contributes to the growing body of zero-inflated models for microbial sequencing data. Much of the existing work focuses on differential abundance analysis and analyzes one taxon at a time. There has been less emphasis on zero-inflated methods that model microbial dependency structures, which is the chief priority of the work herewith.

An important first step in many microbiome studies is to identify possible distinct microbial communities in a given data set and to identify the important bacterial taxa that characterize these communities. The observed data from typical microbiome studies are high dimensional count data with excessive zeros due to both absence of species (structural zeros) and low sequencing depth or dropout (sampling zeros). Although methods have been developed for identifying microbial communities using mixture models of counts, these methods do not account for the excessive zeros of the data and do not differentiate structural from sampling zeros. In Chapter 2, a zero-inflated Latent Dirichlet Allocation model (zinLDA) for the sparse count data seen in microbiome studies is introduced (Deek and Li, 2020). zinLDA builds on the flexible Latent Dirichlet Allocation model of Blei et al. (2003) and allows for zero-inflation in observed counts. This work develops an efficient Markov chain Monte Carlo (MCMC) sampling procedure to fit the model. Results from simulations show zinLDA provides better fit to the data and is able to separate structural zeros from sampling zeros. Furthermore, zinLDA is applied to a data set from the American Gut Project and identifies microbial communities characterized by different bacterial genera.

Moreover, covariation networks among bacterial taxa in microbial communities provide important

insights into the connectivity and structure of the ecosystem. Though, quantification of microbial covariations from 16S rRNA and shotgun metagenomic sequencing data is difficult due to their zero enriched nature. Existing methods using simple correlations or log-based transformations cannot capture the covariations observed in real data sets. Accordingly, this dissertation details two novel copula models with mixture margins to estimate the covariation between a pair of zero-inflated variables.

First, copula models with zero-beta mixture margins are proposed in Chapter 3 for the estimation of taxon-taxon covariations using normalized microbial relative abundance data from cross-sectional studies. Copulas allow for separate modeling of the dependence structure from the margins, easy adjustment of observed confounders in the marginal distributions, and uncertainty measurement. It is shown that a two-stage maximum likelihood approach provides accurate estimation of model parameters. A corresponding two-stage likelihood ratio test for the copula dependence parameter is derived and is used for constructing covariation networks. Simulation studies show that the test is valid, robust, and more powerful than tests based upon Pearson's and rank correlations. Furthermore, data from the American Gut Project is used again to demonstrate that the method can be applied to build biologically meaningful microbial networks.

Second, these models can be extended and applied to longitudinal microbiome data to construct temporally conserved covariation networks. Longitudinal microbiome studies, in which data on a single subject is collected repeatedly over time, are becoming increasingly common in biomedical research. Such studies provide an opportunity to study the inherently dynamic nature of a microbiome in a way that cannot be done using cross-sectional studies. In Chapter 4, a random-effects copula models with zero-beta mixture margins is developed to identify biologically meaningful temporally conserved covariation between two bacterial taxa, while accounting for the excessive zeros seen in 16S rRNA and metagenomic sequencing data. Estimating conserved covariation is of interest as a measure of biological and ecological system robustness. The method assumes a random-effects model for the dependence parameter in the copulas, which captures the conserved microbial covariation while allowing for time-specific dependence parameters. A Monte Carlo EM algorithm is developed

for efficient estimation of model parameters and a corresponding Monte Carlo likelihood ratio test for the mean conserved dependence parameter of the random effect. Simulation studies show that the proposed method provides an empirically unbiased estimate of the mean dependence parameter and the proposed test controls the Type I error rate better than naive methods. Analysis of the longitudinal pediatric DIABIMMUNE cohort identifies changes in both local and global patterns of conserved microbial covariation networks in infants treated with antibiotics. Furthermore, results show that the no antibiotics network is less dependent on individual taxon, thus making it more stable than the antibiotics network and less fragile when exposed to network attacks. This dissertation concludes in Chapter 5, with a discussion of its key findings and contributions to dependency modeling in metagenomic sequencing studies. Further directions and applications of this work to multiomics data integration and single-cell genomics are also introduced.

## CHAPTER 2

### A ZERO-INFLATED LATENT DIRICHLET ALLOCATION MODEL FOR MICROBIOME STUDIES

#### 2.1. Introduction

A significant portion of statistical methodology for the analysis of microbiome data has focused on its high-dimensional nature. A single 16S rRNA gene sequencing sample can produce tens of thousands of sequencing reads from hundreds of different amplicon sequence variants (ASVs). Due to this high dimensional property, of particular interest are techniques for dimensionality reduction. Commonly used methods include principal coordinate analysis (PCoA) with distance measures, such as weighted and unweighted UniFrac distance and Bray-Curtis dissimilarity, or canonical correlation analysis with sparsity assumptions (Hawinkel et al., 2019; Chen et al., 2013). More recently, studies have begun to focus on understanding microbial dynamics within the human microbiome. Single species analyses, that focus on one species at a time in a “parts-list” fashion, are not able to capture complex and dynamic interactions. These interspecies interactions form the basis of distinct underlying subcommunity structures and failing to account for them contributes to the data heterogeneity commonly seen in microbiome studies. As such, network-based approaches have been successfully applied in this area (Faust and Raes, 2012; Layeghifard et al., 2017). These methods use co-occurrence or correlation measures to identify pairwise interactions in cross-sectional studies (Faust et al., 2012; Friedman and Alm, 2012; Kurtz et al., 2015). Others use temporally conserved covariance to identify interactions in longitudinal studies (Raman et al., 2019).

Generative probabilistic mixture models are able to act as a dimensionality reduction technique while simultaneously describing microbial dynamics via subcommunity identification. When applied to microbiome data the latent variable(s) in a mixture model have meaningful biological connotations. Specifically, they represent distinct subcommunity profiles, or structures, that give rise to the observed samples. The simplest of these is the Dirichlet-multinomial mixture model (Holmes et al., 2012). This model is a generalization of the Dirichlet-multinomial hierarchical model. Rather

than assuming that all samples in a cohort are generated from a single community profile, as the Dirichlet-multinomial model does, the mixture model assumes the cohort contains many different subcommunity structures and each of the samples is generated by one of them (Holmes et al., 2012). As such, a sample can be described by its subcommunity assignment rather than a high-dimensional vector of ASV counts. Though, the Dirichlet-multinomial mixture model may still be too restrictive to accurately capture microbial community structures and all the heterogeneity seen in microbiome studies (Sankaran and Holmes, 2019). It is biologically plausible that an individual’s microbiome is comprised of numerous subcommunities, rather than just one, mixing together to varying degrees. The Latent Dirichlet Allocation (LDA) model describes such a generative process (Blei et al., 2003). Samples are defined by their mixture probabilities for each of the subcommunities rather than belonging to a single subcommunity. Technically speaking, LDA differs from the Dirichlet-multinomial mixture model by sampling the latent community variable repeatedly within a sample, once per sequencing read, rather than just once for the entire sample (Blei et al., 2003; Griffiths and Steyvers, 2004).

Latent Dirichlet Allocation has been successful in identifying functional subcommunities of microbes in the human gut and the soil of tropical forests (Sankaran and Holmes, 2019; Hosoda et al., 2020; Sommeria-Klein et al., 2020; Higashi et al., 2018). Despite this, it has been noted that LDA is prone to over-smoothing of microbial counts, which are known to be sparse (Sankaran and Holmes, 2019). This can be attributed to the Dirichlet distribution being insufficient to capture the overdispersion and zero-inflation of microbiome data. The distribution only has one dispersion parameter and inherently imposes a negative correlation between component counts, which may lead to spurious associations (Tang and Chen, 2019). Moreover, the model assumes that each species has a non-negative probability of belonging to every subcommunity. This implies that all species contribute to every subcommunity, even if only with low probability. However, it is more likely that the presence of one species in a community prevents the presence of another.

As such, it would be advantageous to be able to identify community structures that are only composed of a subset of microbial species present in a data set. Thus estimating some of the taxa

membership probabilities for each subcommunity to be zero. We propose a zero-inflated Latent Dirichlet Allocation (zinLDA) model that is flexible enough to capture sparse subcommunities of microbiota. In the following section we detail the generative process of the LDA model and our zero-inflated LDA model. We also provide information on how to estimate model parameters using Markov chain Monte Carlo (MCMC) methods. We apply both models to simulation studies and real data using data from the American Gut Project to directly compare the two and highlight how our proposed method provides better fit to microbiome data.

## 2.2. Latent Variable Mixture Modeling for Microbiome Studies

### 2.2.1. Notation and terminology

Data in microbiome studies often comes from high-throughput sequencing of the 16S rRNA gene. A single biological sample can be represented by a vector of taxon counts with each component representing the number of reads aligned to that specific classification (e.g. ASV, species, genus). The following definitions and notations will be of help in defining a generative probabilistic model for microbiome studies:

- $w_{dn}$  is the  $n^{th}$  observed sequencing read in the  $d^{th}$  biological sample. Sequencing reads are represented by  $V$ -length vectors with a single non-zero component whose value is equal to one, where  $V$  is the number of unique taxa in the study.
- $w_{dn}^i$  represents that the  $n^{th}$  sequencing read in the  $d^{th}$  sample belongs to the  $i^{th}$  unique taxa ( $i = 1, \dots, V$ ).
- $\mathbf{w}_d = (w_{d1}, \dots, w_{dN})$  is the  $d^{th}$  biological sample consisting of  $N$  sequencing reads.
- A cohort  $\mathbf{D} = (\mathbf{w}_1, \dots, \mathbf{w}_D)$  is a collection of all biological samples in the study.

### 2.2.2. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a probabilistic model that is flexible enough to describe the generative process for discrete data in a variety of fields from text analysis to bioinformatics (Blei et al., 2003). When applied to microbiome studies, LDA provides the following generative process for the taxon

counts in a cohort  $\mathbf{D}$ .

---

**Algorithm 1** Generative process for Latent Dirichlet Allocation.

---

- 1: For each of the  $K$  subcommunities, indexed by  $j$ :
  - 2:   Choose  $\beta^{(j)} \sim \text{Dir}(\eta)$ .
  - 3: For each biological sample  $\mathbf{w}_d$  in the cohort:
  - 4:   Choose  $\theta^{(d)} \sim \text{Dir}(\alpha)$ .
  - 5: For each of the  $N$  sequencing reads,  $w_{dn}$ :
  - 6:   Choose a subcommunity  $z_{dn} \sim \text{Multinomial}(1, \theta^{(d)})$ .
  - 7:   Choose a taxon  $w_{dn}$  from  $P(w_{dn}|z_{dn}, \beta)$ , a multinomial probability distribution conditional on the subcommunity  $z_{dn}$ .
- 

Figure 2.1 provides a graphical model representation of LDA. In this model,  $\beta = [\beta_{ij}]$  fully describes the taxa distribution for each subcommunity. The probability that the  $i^{\text{th}}$  taxa belongs to the  $j^{\text{th}}$  subcommunity is denoted by  $\beta_{ij}$ . Note that the taxa distribution is cohort-specific meaning that it is common across all samples and is only estimated once per cohort. The mixture probabilities for the subcommunities of the  $d^{\text{th}}$  sample are denoted by a  $K$ -length vector,  $\theta^{(d)}$ , with  $\theta_{dj}$  representing the mixture probability of the  $j^{\text{th}}$  subcommunity in the  $d^{\text{th}}$  sample. Here,  $K$  is the number of underlying subcommunities and is assumed to be known a priori. Additionally,  $z_{dn}$  is the subcommunity assignment for sequencing read  $w_{dn}$ . Both hyperparameters  $\eta$  and  $\alpha$  are assumed to be symmetric and are defined once for the whole cohort.

Intuitively,  $\beta_{ij} = P(w_{dn}^i | z_{dn} = j)$  determines which taxa are important to subcommunity  $j$  and  $\theta_{dj} = P(z_{dn} = j)$  determines which subcommunities are important in the  $d^{\text{th}}$  sample. Moreover, the LDA model acts as a “soft” clustering technique by allowing samples to be composed of multiple subcommunities. Geometrically, the parameter space of  $\beta$  and  $\theta$  can be thought of in terms of a simplex space. The taxa per subcommunity distribution belongs the  $V-1$  simplex, such that  $\beta^{(j)} \in S^{V-1}$ . Meanwhile,  $\theta^{(d)}$ , the subcommunity distribution per sample can be represented by a randomly selected point in the  $(K-1)$ -dimensional simplex,  $S^{K-1}$ . This is different from the Dirichlet-multinomial mixture model in which  $\theta^{(d)} = \theta$  is assumed to be fixed across all samples and can be represented by the vertices of  $S^{K-1}$ .

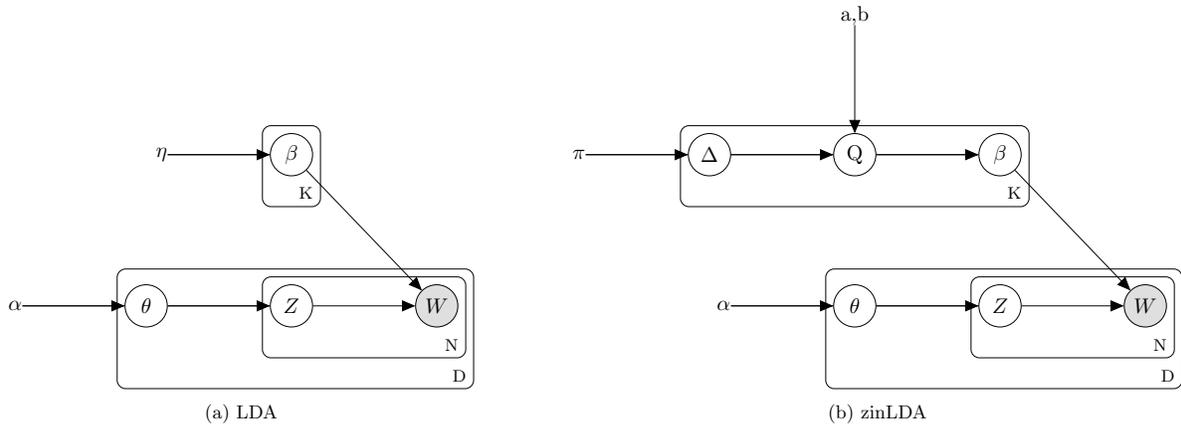


Figure 2.1: Plate diagrams of (a) Latent Dirichlet Allocation and (b) zero-inflated Latent Dirichlet Allocation. Nodes represent parameters and random variables, shading denotes the observed data, and boxes represent repeated sampling. The outer box is denoted with  $D$  for once per biological sample, the inner box with  $N$  for sampling per sequencing read, and the upper box with  $K$  for per subcommunity.

### 2.2.3. Zero-inflated Latent Dirichlet Allocation (zinLDA)

We propose a modification to the Latent Dirichlet Allocation model that allows the latent subcommunity organization to be composed of both structural zeros, taxa that truly do not belong to the community, and sampling zeros, taxa that belong to the community but are not captured due to low sequencing depth or dropout. Understanding and identifying structural zeros in the data is biologically interesting as it provides insights into the absence of certain taxa in a given community.

The zero-inflated generalized Dirichlet (ZIGD) distribution is able to model both sources of zeros (Tang and Chen, 2019). The generalized Dirichlet (GD) distribution is an extension of the Dirichlet that allows for a more flexible covariance structure via the introduction of additional parameters (Connor and Mosimann, 1969). Though, it should be noted that the GD distribution alone does not model structural zeros. To do so, we must modify the unique relationship between the GD distribution and a set of mutually independent beta random variables. By adding a zero-inflation probability,  $\pi$ , to each of the beta random variables we arrive at the zero-inflated generalized Dirichlet distribution. Formally, a length- $V$  vector of ZIGD compositions, denoted by  $\beta = \{\beta_1, \dots, \beta_V\}$ ,

can be formulated from a set of mutually independent zero-inflated beta random variables, which we denote by  $\mathbf{Q} = \{Q_1, \dots, Q_{V-1}\}$ , with zero-inflation probabilities,  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_{V-1}\}$  and parameters  $(a, b)$  of the beta distributions. The relationship between the two random variables can be described as follows:  $\beta_1 = Q_1$ ,  $\beta_l = \prod_{i=1}^{l-1} (1 - Q_i)$  for  $l = 2, \dots, V-1$ , and  $\beta_V = 1 - \sum_{i=1}^{V-1} \beta_i$  (Tang and Chen, 2019). Furthermore, we introduce an indicator variable,  $\Delta_i = I(\beta_i = 0) = I(Q_i = 0)$ , to identify structural zeros. For subcommunity  $j$ , let there be  $L_j$  taxa with  $\beta_{ij} > 0 \Leftrightarrow \Delta_{ij} = 0$ . Then let  $\mathbf{U}_j$  denote the set of indices of the non-zero taxa probabilities for subcommunity  $j$ ,  $\mathbf{U}_j = \{u_{1j}, \dots, u_{L_jj}\}$ , and  $\bar{\mathbf{U}}_j$  be its complement.

Replacing the Dirichlet( $\eta$ ) prior on  $\boldsymbol{\beta}$  with a ZIGD( $\pi, a, b$ ) gives a zero-inflated Latent Dirichlet Allocation (zinLDA) model. The zinLDA model assumes the following generative process for a cohort  $\mathbf{D}$ :

---

**Algorithm 2** Generative process for zero-inflated Latent Dirichlet Allocation.

---

- 1: For each of the  $K$  subcommunities, indexed by  $j$ :
  - 2:   Choose  $\boldsymbol{\Delta}^{(j)} \sim \text{Ber}(\boldsymbol{\pi})$ .
  - 3:   Choose  $\boldsymbol{\beta}^{(j)} \sim \text{ZIGD}(\boldsymbol{\pi}, a, b)$ .
  - 4: For each biological sample  $\mathbf{w}_d$  in the cohort:
  - 5:   Choose  $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\boldsymbol{\alpha})$ .
  - 6: For each of the  $N$  sequencing reads,  $w_{dn}$ :
  - 7:   Choose a subcommunity  $z_{dn} \sim \text{Multinomial}(1, \boldsymbol{\theta}^{(d)})$ .
  - 8:   Choose a taxon  $w_{dn}$  from  $P(w_{dn}|z_{dn}, \boldsymbol{\beta})$ , a multinomial probability distribution conditional on the subcommunity  $z_{dn}$ .
- 

In this model we assume hyperparameters  $\pi$ ,  $a$ ,  $b$ , and  $\alpha$  are symmetric and are defined once for the whole cohort. Comparing the graphical model representation of zinLDA to that of the LDA model (Figure 2.1) underscores the differences between the two, particularly with respect to modeling  $\boldsymbol{\beta}$ .

### 2.3. A Collapsed Gibbs Sampler for Model Inference

We adopt a Bayesian framework for parameter estimation and inference. As such, inference for the zinLDA model is centered around the posterior distribution:

$$P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta} | \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\pi}, a, b) = \frac{P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\pi}, a, b)}{P(\mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\pi}, a, b)}. \quad (2.1)$$

Calculation of this distribution cannot be done directly because the marginalization required to find the normalizing constant,  $P(\mathbf{w}|\alpha, \pi, a, b)$ , is intractable. As such, approximate methods are necessary for parameter estimation. Variational inference may be used to find parameter estimates by maximizing an approximation to the true posterior. Alternatively, a Markov chain Monte Carlo procedure, such as Gibbs sampling, may be used to generate samples from the target posterior distribution for inference. It is worthy to note that due to the fact that both the Dirichlet and ZIGD distributions are conjugate prior for the multinomial distribution, using a collapsed Gibbs sampler that marginalizes over  $\beta$  and  $\theta$  gives a tractable solution, even more so than had collapsing not been performed. For this reason, we propose a collapsed Gibbs sampler for the joint posterior distribution of  $\mathbf{z}$  and  $\Delta$ ,  $P(\mathbf{z}, \Delta|\mathbf{w})$ , where:

$$P(\mathbf{z}, \Delta|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z}, \Delta)}{P(\mathbf{w})} = \frac{P(\mathbf{w}|\mathbf{z}, \Delta)P(\mathbf{z})P(\Delta|\pi)}{\sum_{\Delta} \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z}, \Delta)} \quad (2.2)$$

Integration over  $\beta$  and  $\theta$  can be done separately as the former only appears in  $P(\mathbf{w}|\mathbf{z}, \beta, \Delta)$  and the latter only in  $P(\mathbf{z}|\theta)$ . In Gibbs sampling, each state of the chain is taken as an assignment of each  $z_{dn}$  and  $\Delta_{ij}$ . These states are sampled conditional on the observed data and all the other parameters in the model at their current state. Thus, to perform the sampling, the full conditional distributions,  $P(z_{dn} = j|\mathbf{w}, \mathbf{z}_{-n}, \Delta)$  and  $P(\Delta_{ij} = 1|\mathbf{w}, \mathbf{z}, \Delta_{-i})$ , must be known. These distributions have closed form solutions due to the conjugate prior property of the Dirichlet and ZIGD distributions and can be found probabilistically (Appendix A)

$$P(z_{dn} = j|\mathbf{z}_{-n}, \mathbf{w}, \Delta) \propto \begin{cases} \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{1j} \\ \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \prod_{t < i, t \in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } u_{1j} < i < u_{Lj} \\ \prod_{t < i, t \in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{Lj} \\ 0 & \text{if } i \notin U_j \end{cases} \quad (2.3)$$

$$P(\Delta_{ij} = 1 | \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z}) = \begin{cases} 0 & \text{if } n_j^{(i)} > 0 \\ \frac{\pi_{ij}}{\pi_{ij} + (1 - \pi_{ij}) \frac{B(a_{ij}^{(z)}, b_{ij}^{(z)})}{B(a, b)}} & \text{if } n_j^{(i)} = 0 \end{cases} \quad (2.4)$$

where  $z_{dn}$  is the subcommunity assignment for sequencing read  $w_{dn}^i$ . We define  $n_{j,-n}^{(i)}$  as the number of times the  $i^{th}$  taxa is assigned to the  $j^{th}$  subcommunity and  $m_{j,-n}^{(d)}$  as the number of times the  $j^{th}$  subcommunity occurs in the  $d^{th}$  sample, both excluding the current subcommunity assignment of  $z_{dn}$ . Additionally, we define  $a_{ij}^{(z)} = a + n_j^{(i)}$  and  $b_{ij}^{(z)} = b + n_j^{(i+1)} + \dots + n_j^{(V-1)}$ .

The chain is initialized with informative values for the  $z_{dn}$  variables by sampling from a multinomial distribution with taxa probabilities equal to the  $\beta_{ij}$  estimates from a standard LDA model. Once the chain has been run long enough to guarantee sufficient convergence, a set of the initial runs is removed as a burn-in period, and the remaining are taken as a set of samples from the target posterior distribution. As such, for each run, we can calculate estimates of  $\beta$  and  $\theta$  as follows using the posterior predictive distribution:

$$\hat{\beta}_{ij} = P(w_{new}^{(i)} | z_{new}^{(i)} = j, \mathbf{w}, \mathbf{z}, \mathbf{\Delta}) = \begin{cases} \frac{a + n_j^{(i)}}{a + n_j^{(i)} + b_{ij}^{(z)}} & \text{if } i = u_{1j} \\ \frac{a + n_j^{(i)}}{a + n_j^{(i)} + b_{ij}^{(z)}} \prod_{t < i, t \in U_j} \frac{b_{tj}^{(z)}}{a + n_j^{(t)} + b_{tj}^{(z)}} & \text{if } u_{1j} < i < u_{Lj} \\ \prod_{t < i, t \in U_j} \frac{b_{tj}^{(z)}}{a + n_j^{(t)} + b_{tj}^{(z)}} & \text{if } i = u_{Lj} \\ 0 & \text{if } i \notin U_j \end{cases} \quad (2.5)$$

$$\hat{\theta}_{dj} = P(z_{new} = j | \mathbf{w}, \mathbf{z}) = \frac{m_j^{(d)} + \alpha}{m^{(d)} + K\alpha} \quad (2.6)$$

The final estimate of  $\theta$  is defined as its posterior mean across all the runs. The final estimate of  $\beta$  can be found in a multi-step process. First, calculate the posterior mean of  $\Delta_{ij}$  across all runs, which is equivalent to a posterior estimate of  $\pi_{ij}$ . Then dichotomize  $\hat{\pi}_{ij}$  according to  $I(\hat{\pi}_{ij} \geq 0.5)$ .

Next, assign  $\hat{\beta}_{ij} = 0$  for any  $I(\hat{\pi}_{ij} \geq 0.5) = 1$ , otherwise assign  $\hat{\beta}_{ij}$  its respective posterior mean and normalize within each subcommunity such that  $\sum_i \beta_{ij} = 1$ .

## 2.4. Simulation Studies

### 2.4.1. Parameter specification

We conducted simulation studies to compare estimation accuracy and model fit between the proposed zinLDA and the standard LDA models. The data was simulated from a true zinLDA model, following the steps specified by the generative algorithm given previously. First, we selected the total number of taxa ( $V$ ) to be 120 across 150 independent microbial samples. Next, the total number of reads in each sample were drawn from a discrete Uniform distribution with a lower bound of 5000 and upper bound of 25000. These parameters were selected to reflect real microbiome data sets aggregated to the genus-level classification. The number of subcommunities ( $K$ ) was selected as five. The hyperparameter  $\alpha$  of the Dirichlet distribution on  $\theta$  was set to  $50/K$ , as suggested for the original LDA model (Griffiths and Steyvers, 2004). Additionally, the hyperparameters  $\pi$ ,  $a$ , and  $b$  of the zero-inflated generalized Dirichlet distribution on  $\beta$  were set to 0.4, 0.05, and 10, respectively. After running the simulation algorithm, the taxa that had a zero count for every sample, meaning a prevalence of 0%, were removed as such taxa would not be observed in a real data analysis. This reduced the total number of observed taxa ( $V_{obs}$ ) to 87.

A zinLDA model with five subcommunities was fit to the simulated data set. Hyperparameters  $\alpha$ ,  $\pi$ ,  $a$ , and  $b$  were set to their true values, as specified under simulation. Likewise, a standard LDA model with five subcommunities was fit, with default hyperparameter values of  $50/K$  and 0.1 for  $\alpha$  and  $\eta$ , respectively, as suggested in Griffiths and Steyvers (2004). To deal with the label switching problem commonly seen in Bayesian inference of mixture models, we used a method previously proposed to compare labels from an LDA model to their ground-truth. The pairwise Pearson correlation was calculated for each true-estimated subcommunity pair. The pair with the highest correlation is matched, then the pair with the next highest correlation among the remaining is matched, and so on until all true-estimated pairs are uniquely matched (Sankaran and Holmes, 2019).

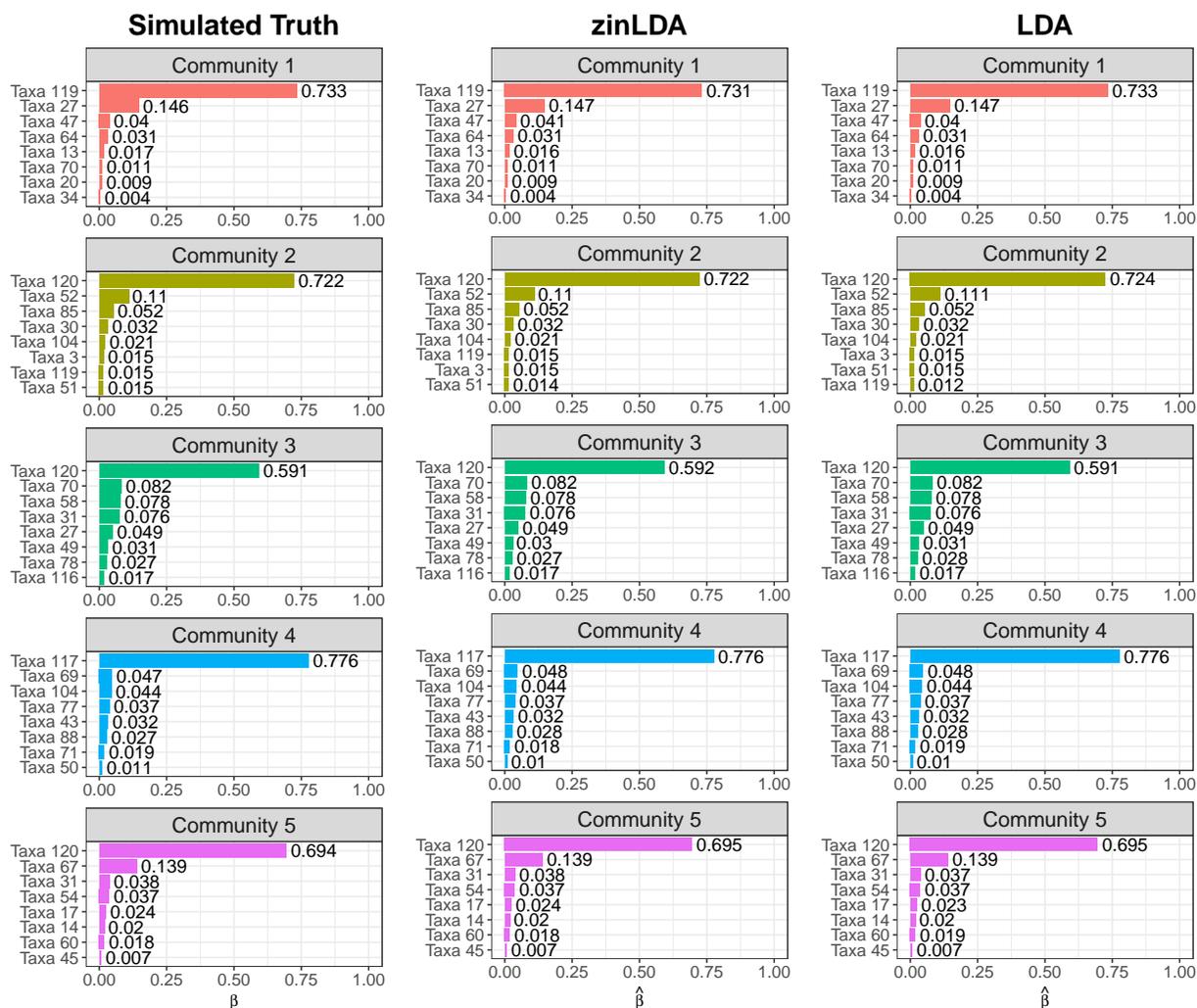


Figure 2.2: Bar graphs of the top eight taxa for each of the five subcommunities alongside their corresponding  $\hat{\beta}_{ij}$  values. The first column contains the ground truth (i.e., top taxa from simulation). The second and third columns are the estimated top taxa from the zinLDA and LDA models, respectively.

### 2.4.2. Model fit and sensitivity of zinLDA

To determine how well zinLDA is able to capture the latent community structure we compared the estimated  $\beta_{ij}$  for the top eight taxa per community to their true value and estimated value from the standard LDA model. Figure 2.2 shows that both zinLDA and LDA correctly identify all of the top microbial taxa for each of the five subcommunities. Moreover, estimates from both models show low bias. We also investigated how misspecification of the number of subcommunities influences zinLDA’s ability to recover the representative taxa. An underspecified model, with one too few communities, collapses the representative taxa of two of the subcommunities together, thus resulting in both upwardly and downwardly biased estimates of  $\beta_{ij}$ . The remaining three subcommunities have their representative taxa recovered and their respective  $\beta_{ij}$  estimates were not effected. Likewise, for an overspecified model, with one too many communities, it is able to accurately detect the five true subcommunity structures, as specified under simulation, but identifies an additional nonsensical subcommunity that is composed of only one taxa (Figure A.1).

Fit of the two models was assessed through posterior predictive checks (Gelman et al., 1996). For each model, the posterior predictive distribution was used to simulate 100 data sets with the same dimensions as the original. The rationale behind using posterior predictive checks to assess model fit is as follows: if the model provides reasonable fit then the data simulated from the posterior predictive distribution, which is conditional on the observed data ( $X_{obs}$ ) and the current model, should “look similar” to the observed data. We quantify how similar the observed data and the posterior predictive simulated data are by the test statistic  $T(X) = X_i$ , the count for the  $i^{th}$  taxa. Figure 2.3 plots the results from the posterior predictive checks. Each panel corresponds to a single biological sample. The  $y$ -axis plots  $T(X)$  on the *asinh* scale. The  $x$ -axis plots each of the 87 taxa, ordered from smallest to largest based on the observed data for that sample. For large taxon counts we see that both models do well, with median values of both being similar to the true observed values. In contrast, we see that for small taxon counts the zinLDA model outperforms LDA. Specifically, for zero counts the zinLDA model is able to accurately estimate these counts better than its LDA counterpart. Across the 50 data sets simulated from the posterior predictive

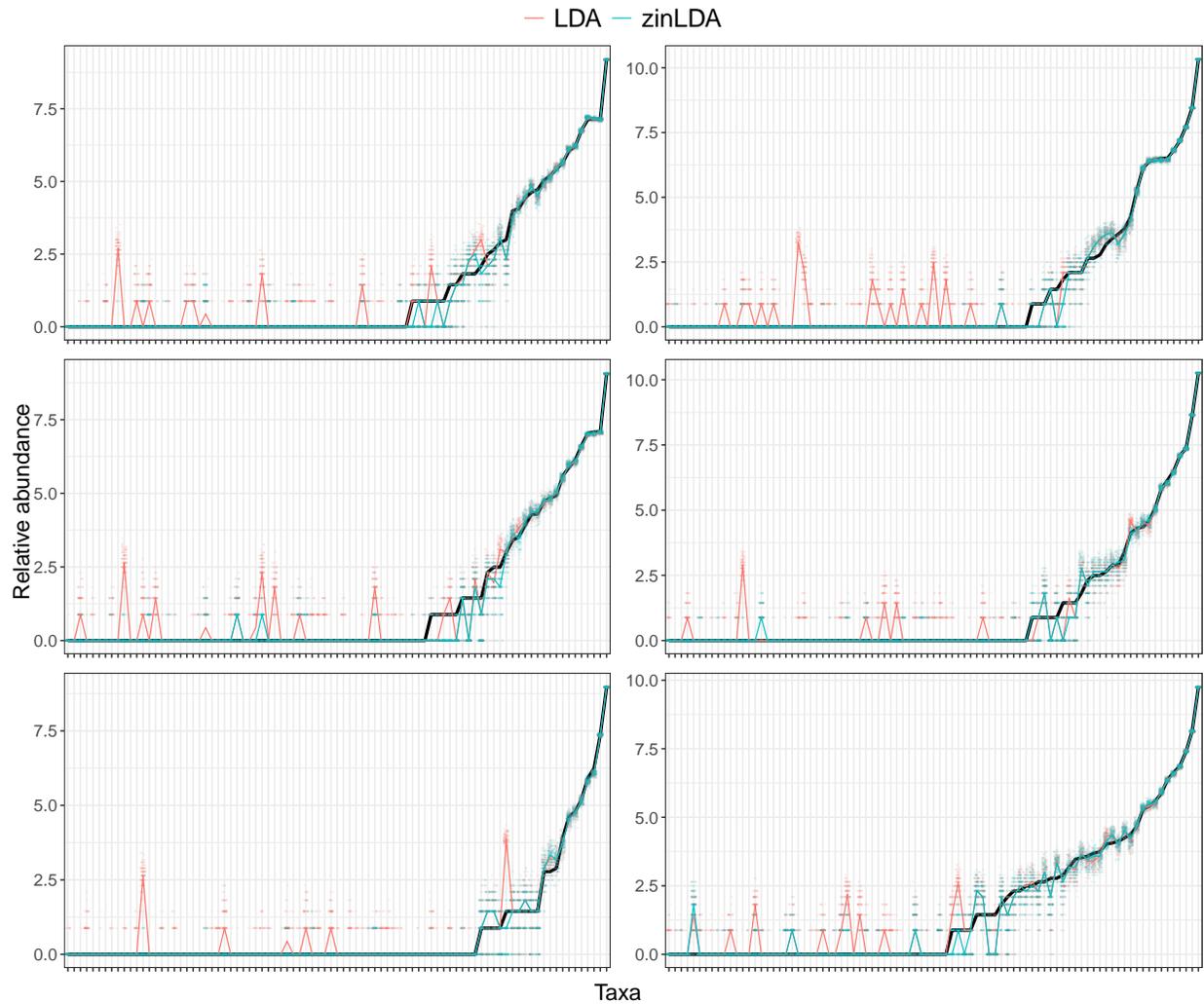


Figure 2.3: Observed and posterior predictive simulated  $asinh$ -transformed taxon counts plotted in order of increasing true abundance. Each panel is a different biological sample. The solid black line represents the true counts. The pink and blue points are the counts from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median counts across the 50 data sets.

distribution, zinLDA exhibits less over-smoothing for small taxon counts compared to the original LDA model. Thus, this is an indication that the zinLDA model provides better fit to the data than LDA.

To quantify how well the zinLDA model is able to distinguish between rare and absent taxa in each subcommunity we calculated its sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We defined a “positive” outcome as being a structural zero,  $\Delta_{ij} = 1$ , and a “negative” outcome as being a non-zero probability of belonging to that subcommunity,  $\Delta_{ij} = 0$ . The results show that under these simulation settings zinLDA can differentiate sampling and structural zeros with reasonable sensitivity and specificity (Table 2.1). Upon further examination the data, we note that the model we used to generate the data resulted in many taxa with very small true non-zero probabilities, making it very difficult to separate sampling zeros from structural zeros. To further demonstrate this point, we ran two additional simulations to see how different model parameters affect the posterior inference of being a structural zero. Both simulations reduced the number of taxa ( $V$ ) to 50, but one also changed hyperparameter  $a$ , of the ZIGD distribution, to 0.5 from 0.05. Table 2.1 shows that reducing the number of taxa without changing the value of  $a$  reduces the model’s ability to differentiate between the two sources of zeros. In contrast, reducing  $V$  and also increasing  $a$  significantly increases the model’s ability to accurately detect structural zeros, with such a model having a sensitivity of 0.9 and PPV of 0.92. The sharp difference in the values of these diagnostic metrics between the models can be attributed to the fact that  $V$ ,  $a$  and  $b$  all influence the  $\beta_{ij}$  values, which in turn influences the probability of observing a sampling zero. For example, decreasing  $V$  without changing  $a$  reduces many of the  $\beta_{ij}$  values, thus increasing the probability of observing a sampling zero. On the other hand, decreasing  $V$  in conjunction with increasing  $a$  increases many of the  $\beta_{ij}$  values and therefore decreases the probability of observing a sampling zero.

## 2.5. Identification of Gut Microbial Subcommunities in a Healthy Cohort

The American Gut Project (AGP) is a self-selected and open platform cohort. Citizen-scientists primarily in the United States, United Kingdom, and Australia, opted into the study, paid a fee to

Parameters	Sensitivity	Specificity	PPV	NPV
$V = 50, a = 0.5$	0.90	0.94	0.92	0.93
$V = 50, a = 0.05$	0.51	0.51	0.40	0.61
$V = 87, a = 0.05$	0.73	0.67	0.59	0.79

Table 2.1: Comparison of estimated structural zero taxa from the zinLDA model to true structural zero taxa from simulation across different parameter settings using sensitivity, specificity, positive predictive value, and negative predictive value. A “positive” results is assumed to be  $\hat{\beta}_{ij} = 0$ .

offset the cost of sample processing and sequencing, and gave informed consent (McDonald et al., 2018). All subjects provided a fecal microbiome sample and self-reported metadata. The sequencing protocol used was identical to that of the Earth Microbiome Project (McDonald et al., 2018; Gilbert et al., 2014). The AGP microbial 16S rRNA gene sequencing data and metadata are publicly available in The European Bioinformatics Institute repository under the accession ERP012803.

This analysis used a prior subset of the AGP data consisting of 3679 subjects. Reads that were ambiguously assigned or unassigned at the genera level were removed. Moreover, genera with a prevalence of less than 20% across all samples were removed. After this filtering of the microbial genera, any samples with a total number of reads of zero were removed. This left 3566 samples and 70 unique genera for downstream analyses. A random subset of 1000 subjects from the AGP data was sampled, a zinLDA model with five subcommunities and hyperparameter values specified the same as in the simulation study was fit. When possible, the choice of the number of latent subcommunities should be informed by biological or clinical reasoning. In the absence of such, data-driven approaches may be used. In particular for the AGP data,  $K$  was determined by comparing the log-likelihood, AIC, and representative taxa across many models, each with a different number of subcommunities, applied to a set of 1000 independently selected subjects. These results were robust across slight changes in the number of subcommunities. The representative taxa from each subcommunity and their membership probability ( $\beta_{ij}$ ) is shown in Figure 2.4. We observe that each subcommunity is characterized by one single dominant taxa, including *Faecalibacterium*, *Prevotella*, *Bacteroides*, *Acinetobacter* and *Akkermansia*.

Model fit was assessed via posterior predictive checks and compared to that of the standard LDA

model. Since current sequencing technology, such as 16S rRNA gene sequencing, can only provide quantification about the relative abundance of microbial compositions model fit was assessed using both the relative abundance and the observed counts (Appendix A.2 and A.3). The two plots exhibit similar patterns, indicating the difficulty in fitting the small counts in the data. Another explanation of observing such similar model fits is that our analysis did not identify structural zeros with strong evidence in our data. Figure A.4 shows the posterior estimates of the probability of being a structural zero for each of the taxa in each subcommunity, indicating relatively weak evidence of structural zeros.

Finally, to determine whether the model is stable, meaning it detects true subcommunity clusters of co-occurring taxa and is not clustering the noise in the observations, we applied an identical zinLDA model to another set of 1000 AGP microbial samples that is independent from the first. The representative taxa from this validation set are compared to that of the first cohort (Figure 2.4). The subcommunities between the two cohorts were matched using pairwise correlations as done in simulations. The average cosine similarity of the matched subcommunities is 0.80. These results show that the subcommunities identified by zinLDA are highly stable and replicable.

## 2.6. Discussion

The microorganisms that compose the human microbiome form subcommunity-like structures via dynamic and complex interactions with one another. Identifying these structures is imperative for a better understanding of how these microbes influence human-host health. We proposed a zero-inflated Latent Dirichlet Allocation model, an extension of the LDA model that amounts to changing the prior distribution on the taxa per subcommunity distribution to a zero-inflated generalized Dirichlet distribution from a Dirichlet. Despite this change our model retains the advantageous conjugate prior property between the ZIGD and multinomial distributions. As such, we are able to implement an efficient Gibbs sampling algorithm with only one additional step compared to that of LDA for parameter estimation.

zinLDA modifies the LDA model proposed by Blei et al. (2003) to allow for subcommunities to be composed of a subset of all the microbes in a cohort of microbial sequencing samples. Mathemat-

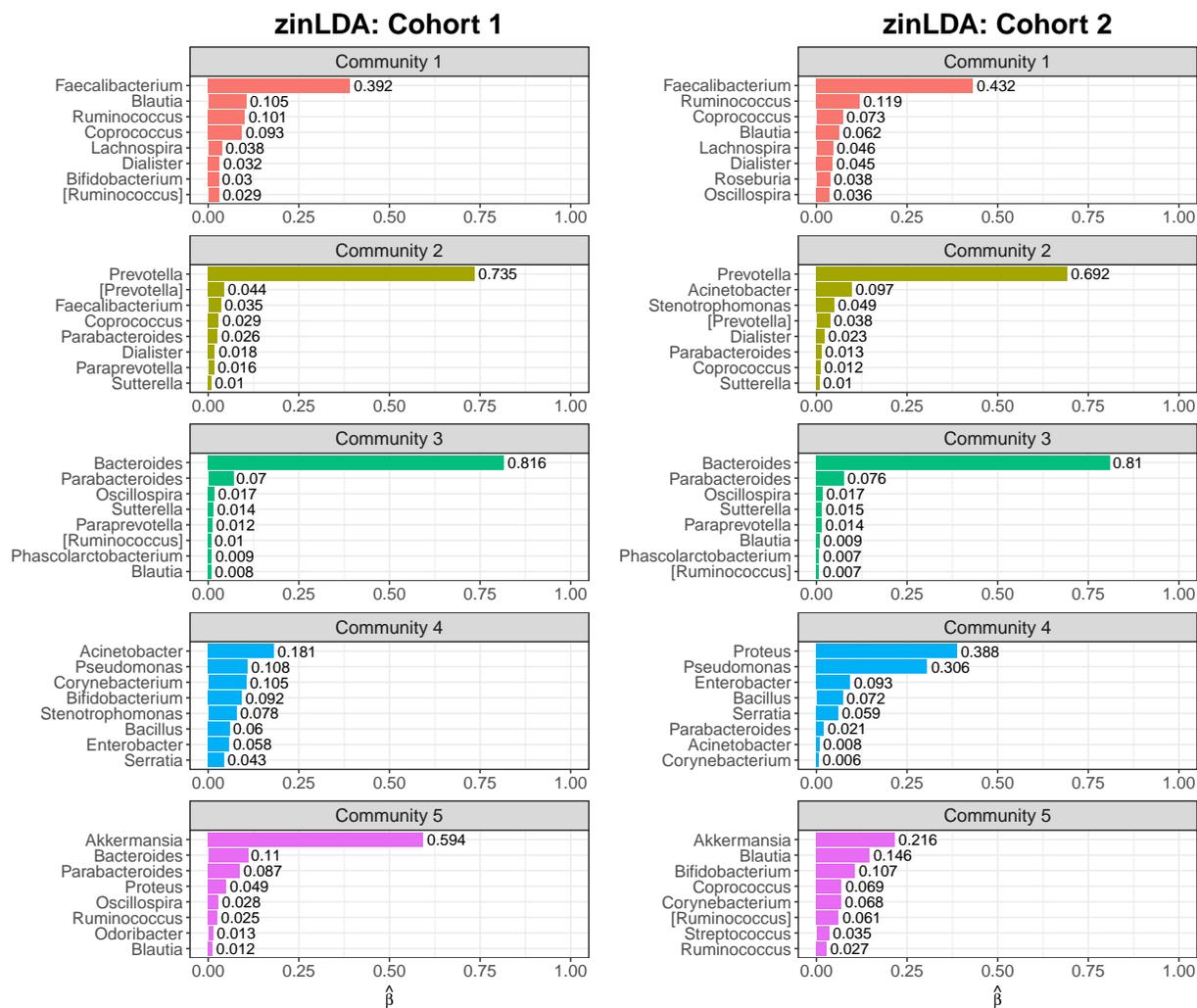


Figure 2.4: Bar graphs of the representative taxa for each of the five subcommunities alongside their corresponding  $\hat{\beta}_{ij}$  values for two independent subsets of 1000 randomly selected subjects from the American Gut Project.

ically, since a subcommunity is defined as a distribution over taxa, this is equivalent to assigning some taxa a zero-probability of belonging to the subcommunity. This is particularly advantageous in microbial analyses as it allows for a clear distinction between sampling and structural zeros within a subcommunity structure. Structural zeros come from those zero-probability taxa; they are truly absent from the community. Sampling zeros come from taxa that do belong to the community, but with low probability, and thus were not captured due to shallow sequencing depth. Due to this adjustment, the zinLDA model can be used to simulate more realistic sparse count data than models such as the Dirichlet-multinomial or Dirichlet-multinomial mixture models.

We used simulation studies to compare the two models and investigate where zinLDA outperforms the standard LDA model. First, we showed that the two performed equally well in identifying the representative taxa for each subcommunity. This is to be expected as the LDA model already does a good job in identifying common taxa and the zinLDA estimates of the community assignment for each sequencing read were initialized using the results from a standard LDA model. The performance gain in using zinLDA is seen when examining the low probability and absent taxa within each subcommunity. The greatest performance gains were made when the probability of being a sampling zero was not too small. Furthermore, we used real data from the citizen scientists of the American Gut Project to show that our method can detect potentially meaningful biological and ecological subcommunities of microbial species. By assigning every sample a probability of belonging to each of these subcommunities we were also able to gather information about population level microbial structures.

As with any Bayesian model, zinLDA requires the hyperparameters to be prespecified. In our analysis of real data sets, we used the same hyperparameters as in our simulations and explored various other choices. For the same number of communities, we observed that the community structures and the representative taxa were not too sensitive to the values of these hyperparameters. Determining the number of clusters or subcommunities is a hard problem, as is the case with any clustering method. For real data analysis, we suggest that the users try different numbers of  $K$ , evaluate the subcommunity structures, and then choose one based on both the sizes of the

communities and also possible biological interpretations. Finally, the zinLDA model can be used to simulate more realistic microbiome count data that allow for both structural and sampling zeros. Such simulations can be used to evaluate various statistical tests developed for microbiome data analysis, including evaluating power of tests for differential abundance and methods for modeling microbiome count data.

## CHAPTER 3

### INFERENCE OF MICROBIAL COVARIATION NETWORKS USING COPULA MODELS WITH MIXTURE MARGINS

#### 3.1. Introduction

While there have been advances in understanding the association between individual microbes and human health, much remains unknown about the relationships between microbes and how they change when comparing diseased and healthy microbial communities. What is known is the microorganisms that compose a microbiome form complex and dynamic interactions not only with their host environment but also with one another (Gerber, 2014; Li, 2015). Such interactions lead to covariations among the bacterial taxa in microbial communities. Most often, ecological patterns of the microbiome are investigated using microbial covariation, including co-occurrence or co-exclusion. The use of marginal, pairwise association measures are ubiquitous in the analysis of high-throughput sequencing data. Covariation analyses are analogous to co-expression analyses, widely used with gene expression data (Zhang and Horvath, 2005). Typically co-occurrence and co-exclusion are defined statistically as a positive or negative correlation, respectively, between two taxa greater than some threshold (Williams et al., 2014; Barberán et al., 2012; Widder et al., 2014).

For microbiome studies, the data are often summarized into a vector of bacterial relative abundances, with excessive zeros. There are some attempts to estimate covariations on the absolute abundance scale, using the relative abundance data, by assuming that such covariations are very sparse (Friedman and Alm, 2012). These methods are all based on log-ratio transformations (Aitchison, 1982; Cao et al., 2019). Despite their ubiquity, such transformations are not well suited for data with excessive zeros, as encountered in microbiome relative abundance data. To perform such transformations, pseudocounts have to be used and their results are difficult to interpret. Even more, the normality assumption of the transformed data often does not hold. Additionally, as shown by Cao et al. (2019), the rate of convergence of such a regularized estimate of the covariance matrix with dimension  $p$  includes the terms  $s_0(p)\{(log p)/n\}$  and  $s_0(p)(s_0(p)/p)$ , which requires that the

sparsity parameter  $s_0(p)$  is very small as  $p$  diverges. Such an assumption may not hold for complex microbial covariation networks.

Alternatively, understanding the conditional independence relationship among all the taxa in a microbial community under a given condition may be of interest. Such relationships can be used to address the question of whether the covariation of two taxa can be explained by other taxa (Kurtz et al., 2015; Yoon et al., 2019). These methods are also based on the log-ratio transformations. Given limited sample sizes and the excessive zeros observed in typical microbiome studies, the multivariate normality assumption does not hold. Further, estimates of such conditional dependence based on regularized estimate of the covariance matrix are very unstable.

Accordingly, due to the reasons outline above, in this chapter we focus on providing a robust and flexible estimate of the covariation of taxon pairs while allowing for covariate adjustments. Instead of attempting to estimate the covariation at the absolute abundance level, which is not possible without strong assumptions, we focus on estimating the covariation at the relative abundance level. Such covariation analyses have appeared widely in microbiome data analysis using relative abundance data and simple correlations (Faust et al., 2012; Faust and Raes, 2012). Due to excessive zeros in the data, commonly used simple sample correlations may lead to loss of power in identifying covarying taxa pairs. To overcome such limitations, we propose a flexible model based procedure to estimate the covariation between any two microbes using their normalized relative abundance. Copula models are particularly well suited for this problem as they allow for separate modeling of the univariate marginal distributions from the dependency structure. Unlike existing methods, copulas also allow for covariate adjustment in the margins and uncertainty quantification of their dependence estimate. We perform estimation on the relative abundance scale by modeling the data using a mixture of zeros and a beta distribution, which has been shown to fit microbiome relative abundance data well (Chen and Li, 2016; Ho et al., 2019). While copulas have been applied to model joint distributions with mixed margins, copula models with both marginal distributions being a mixture of discrete and continuous distributions have not been studied extensively and are the main methodological focus of this chapter.

## 3.2. Copula Models with Mixture Margin Distributions

### 3.2.1. Zero-inflated beta marginal distribution and the copula model

Consider a single microbial sample which can be summarized by the normalized relative abundances of the  $m$ -microbes, denoted by  $(x_1, \dots, x_m) \in [0, 1]^m$ . We assume that each  $x_j$  follows a zero-inflated beta distribution. Accordingly, the marginal density of any given  $x_j$  can be written as,

$$f(x_j) = p_j I_j + (1 - p_j) f_{beta}(x_j | \mu_j, \phi_j) (1 - I_j), \quad (3.1)$$

where we define  $p_j = \Pr(x_j = 0)$ ,  $I_j = I(x_j = 0)$ , and

$$f_{beta}(x_j | \mu_j, \phi_j) = \frac{\Gamma(\phi_j)}{\Gamma(\mu_j \phi_j) \Gamma((1 - \mu_j) \phi_j)} x_j^{\mu_j \phi_j - 1} (1 - x_j)^{(1 - \mu_j) \phi_j - 1}$$

the density function of a beta random variable indexed by mean parameter  $\mu_j$  and dispersion parameter  $\phi_j$ .

It is often of interest to understand the relationship between any pair of microbes, but calculating the joint distribution of a set of non-normal random variables can be tedious and contain many parameters. As such, we propose a copula based approach. Mathematically, a copula is the joint distribution function of a set of uniform random variables,  $\mathbf{A} = (a_1, \dots, a_m)$ . Though, in practice, copulas can be used to describe the distribution function of any set of random variables, such that  $a_k = F_k(x_k)$ , where  $F_k$  is the marginal cumulative distribution function of the  $k^{th}$  variable,  $x_k$ . This is proven by Sklar's theorem, which states that any multivariate joint distribution can be described by two parts: (1) the copula function  $C(\cdot | \theta)$  and (2) the univariate marginal distribution functions  $F_k$  (Sklar, 1959). Therefore, for any pair of microbes we can write the bivariate cumulative distribution of their normalized relative abundances as

$$F(x_i, x_j | \gamma_i, \gamma_j, \theta_{ij}) = C(F_i(x_i; \gamma_i), F_j(x_j; \gamma_j) | \theta_{ij}) = C(u, v | \theta_{ij}), \quad (3.2)$$

where  $U = F_i(\cdot; \gamma_i)$  and  $V = F_j(\cdot; \gamma_j)$  are the univariate zero-inflated beta margins of  $X_i$  and  $X_j$ , respectively, each with parameters  $\gamma = (p, \mu, \phi)^\top$  and  $C(\cdot|\theta)$  is a family of copula functions with dependence parameter  $\theta$ . The copula function links, or ties, together the margins to form the joint distribution. An advantageous property of copulas is that they completely describe the dependency between the margins via their parameter  $\theta$ , thus allowing for separate modeling of the margins and dependence structure.

Moreover, we can specify a set of demographic and clinical variables that affect each microbe's presence-absence probability, mean abundance, and dispersion using generalized linear models. It is for this reason that we used the alternate parameterization of the beta distribution. We assume that parameters of each margin,  $p_k$ ,  $\mu_k$ , and  $\phi_k$  ( $k = i, j$ ) can be specified according to a general class of zero-inflated beta regression models as follows (Ospina and Ferrari, 2012):

$$h_1(p_k) = f_1(\mathbf{Q}_k, \boldsymbol{\rho}_k); \quad h_2(\mu_k) = f_2(\mathbf{W}_k, \boldsymbol{\delta}_k); \quad h_3(\phi_k) = f_3(\mathbf{Z}_k, \boldsymbol{\kappa}_k).$$

We define  $\mathbf{Q}_k$ ,  $\mathbf{W}_k$ , and  $\mathbf{Z}_k$  as the matrix of covariates of interest for the presence-absence probability, mean abundance, and dispersion of the  $k^{th}$  margin, respectively;  $\boldsymbol{\rho}_k$ ,  $\boldsymbol{\delta}_k$ , and  $\boldsymbol{\kappa}_k$  as their corresponding vector of regression parameters; and  $f_1$ ,  $f_2$  and  $f_3$  as some functions of the covariates and regression parameters. As with all GLMs,  $h_1, h_2 : (0, 1) \rightarrow \mathbb{R}$  and  $h_3 : (0, \infty) \rightarrow \mathbb{R}$  are strictly monotonic, twice differentiable link functions. Common choices of link function for  $h_1$  and  $h_2$  are the logit, probit, and log-log. Likewise, the log and square-root link functions are common choices for  $h_3$ .

### 3.2.2. Joint density function of bivariate copula model with two mixture margins

For absolutely continuous margins, the copula distribution function is unique. The joint density function of  $X_i$  and  $X_j$  can be found by taking mixed partial derivatives of the copula function with respect to  $U$  and  $V$ , resulting in  $f(x_i, x_j) = c(u, v|\theta_{ij})f_i(x_i)f_j(x_j)$  where  $c$  is the copula density of  $C$  and  $f_i, f_j$  are the marginal densities of  $x_i$  and  $x_j$ , respectively. For discrete or mixture margins,  $C$  is not unique and the calculation of the joint density function is not as straightforward.

Gunawan et al. (2020) outlines a method for defining the joint density when the margins may belong to any of the three following categories: absolutely continuous, discrete, and mixtures of absolutely continuous and discrete random variables. As such, we can use this general framework to explicitly define the joint density of two zero-inflated beta random variables and use the same notation for consistency. Let  $\mathcal{M} = \{i, j\}$  be the index set,  $\mathcal{C}(\mathbf{x})$  contain the indices of  $\mathbf{x} = \{x_i, x_j\}$  with continuous  $F$  at  $x$ , and  $\mathcal{D}(\mathbf{x}) = \mathcal{M} - \mathcal{C}(\mathbf{x})$  to be the set of indices of  $\mathbf{x}$  for which  $F$  has a jump point at  $x$ . Therefore,  $\mathcal{D}$  is the null set if and only if  $x_i > 0$  and  $x_j > 0$ . Using these two sets Gunawan et al. (2020) defines the joint density of  $x_i$  and  $x_j$  as:

$$f(x_i, x_j) = c_{\mathcal{C}(\mathbf{x})}(\mathbf{b}_{\mathcal{C}(\mathbf{x})}) \prod_{k \in \mathcal{C}(\mathbf{x})} f_k(x_k) \Delta_{\mathbf{a}_{\mathcal{D}(\mathbf{x})}}^{\mathbf{b}_{\mathcal{D}(\mathbf{x})}} C_{\mathcal{D}(\mathbf{x})|\mathcal{C}(\mathbf{x})}(\cdot | \mathbf{b}_{\mathcal{C}(\mathbf{x})}) \quad (3.3)$$

Where  $\mathbf{a} = (F_i(x_i^-), F_j(x_j^-))$  is a vector of cumulative distribution probabilities just before  $x_i$  and  $x_j$  and  $\mathbf{b} = (F_i(x_i), F_j(x_j))$ . Note that when  $x_k > 0$ ,  $F_k(x_k^-) = F_k(x_k)$ . Moreover,  $C_{\mathcal{D}(\mathbf{x})|\mathcal{C}(\mathbf{x})}$  is the copula conditional distribution function of the point masses at zero conditional on the continuous beta part and  $\Delta_{\mathbf{a}}^{\mathbf{b}} g(\cdot) = \Delta_{a_i}^{b_i} \Delta_{a_j}^{b_j} g(\cdot) = g(b_i, b_j) - g(b_i, a_j) - g(a_i, b_j) + g(a_i, a_j)$ . For the bivariate case this implies there are four specifications of the joint density (Appendix B).

The above joint distribution of  $x_i$  and  $x_j$  holds for any choice of copula function  $C$ . Although, in this chapter, we chose to focus on only the Frank copula, whose properties are well suited for microbial covariations. In particular, the Frank copula can model the maximal range of dependence, meaning  $\theta \in \{-\infty, \infty\} \setminus 0$ , with  $\pm\infty$  corresponding to the Fréchet upper and lower bounds. This is particularly advantageous since other Archimedean copulas, such as the Gumbel and Joe copulas, do not permit negative dependence structures, which are likely to be seen in microbial covariations. Also, the magnitude of dependence is symmetric for positive and negative dependencies, including in the tails of the distribution. We use  $C_{Fr}(u, v)$ ,  $C_{v|u, Fr}(u, v)$  and  $c_{Fr}(u, v)$  to denote the Frank copula distribution function, conditional distribution function, and joint density, respectively, where

$$C_{Fr}(u, v) = -\frac{1}{\theta} \log \left\{ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right\}, \quad (3.4)$$

and  $C_{v|u,Fr}(u, v)$  and  $c_{Fr}(u, v)$  can be derived.

Henceforth we assume all copulas are referring to the Frank copula. Now that we have defined the bivariate density of  $x_i$  and  $x_j$ , we can define the likelihood function and use a maximum likelihood estimation procedure for model parameters,  $p_i, \mu_i, \phi_i, p_j, \mu_j, \phi_j$ , and  $\theta_{ij}$ . In the simplest case, of no covariate adjustment, using the typical full maximum likelihood estimation requires a seven dimensional optimization procedure. The numerical optimization of one function with many parameters is more difficult and computationally intensive than the numerical optimization of several functions with fewer parameters. As such, we use a two-stage, or inference-for-margins, procedure that breaks the parameter estimation into several smaller estimation problems (Shih and Louis, 1995; Joe and Xu, 1996).

### 3.3. A Two-Stage Estimation Method and Statistical Inference

#### 3.3.1. A two-stage estimation method

For a sample of size  $n$ , with observed random vectors  $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^2$  that represent the relative abundances of a pair of bacteria  $(i, j)$ , we consider the univariate log-likelihood functions of the zero-inflated beta margins:

$$\ell_k(\gamma_k) = \sum_{l=1}^n \log f_k(x_{lk}|\gamma_k), \quad k \in \{i, j\}$$

and the log-likelihood function for the joint distribution,

$$\ell(\theta, \gamma_i, \gamma_j) = \sum_{l=1}^n \log f(\mathbf{X}_l|\gamma_i, \gamma_j, \theta).$$

Note that we have here, and henceforth will, suppress the subscript on  $\theta$ , implying that we are referring to a given  $(i, j)$  pair of microbes, unless otherwise noted. The two-stage estimation procedure (Shih and Louis, 1995; Joe and Xu, 1996) can be summarized as follows:

1. The log-likelihoods,  $\ell_i$  and  $\ell_j$ , of the two univariate margins are separately maximized to get estimates of their parameters  $\tilde{\gamma}_i$  and  $\tilde{\gamma}_j$ , respectively.

2. The function  $\ell(\theta, \tilde{\gamma}_i, \tilde{\gamma}_j)$  is maximized over  $\theta$  to get  $\tilde{\theta}$ .

We denote  $\boldsymbol{\eta} = (\gamma_i, \gamma_j, \theta)$  as the vector of all parameters,  $\tilde{\boldsymbol{\eta}} = (\tilde{\gamma}_i, \tilde{\gamma}_j, \tilde{\theta})$  as the vector of two-stage estimators, and  $\hat{\boldsymbol{\eta}} = (\hat{\gamma}_i, \hat{\gamma}_j, \hat{\theta})$  as the MLEs that simultaneously maximize the full log-likelihood function.

We begin with the two-stage MLEs of the  $k^{\text{th}}$  zero-inflated beta margin with log-likelihood  $\ell_k$  equal to:

$$\begin{aligned} \ell_k(\boldsymbol{\gamma}_k) &= z_k \log(p_k) + (n - z_k) \log(1 - p_k) + (n - z_k) \log \Gamma(\phi_k) \\ &\quad - (n - z_k) \log \Gamma(\mu_k \phi_k) - (n - z_k) \log \Gamma((1 - \mu_k) \phi_k) \\ &\quad + (\mu_k \phi_k - 1) \sum_{l=1}^n \log(x_{lk}) + ((1 - \mu_k) \phi_k - 1) \sum_{l=1}^n \log(1 - x_{lk}) \end{aligned} \quad (3.5)$$

Where  $z_k = \sum_{l=1}^n I_{lk} = \sum_{l=1}^n I(x_{lk} = 0)$  is the number of observations with  $x_k = 0$  and  $\Gamma(W + 1) = W!$  is the gamma function. We use the Newton-Raphson algorithm to numerically find the MLEs of  $\boldsymbol{\rho}_k$ ,  $\boldsymbol{\delta}_k$ , and  $\boldsymbol{\kappa}_k$ , the GLM regression coefficients.

Now that the marginal two-stage MLEs,  $\tilde{\boldsymbol{\gamma}}_k$ , have been defined they can be plugged into the full likelihood  $\ell(\theta, \gamma_i, \gamma_j)$  to give:

$$\begin{aligned} \ell(\theta, \tilde{\gamma}_i, \tilde{\gamma}_j) &\propto \sum_{l \in S1} \log\{-\theta(e^{-\theta} - 1)\} + \sum_{l \in S1} \log\{e^{-\theta(\tilde{u}_l + \tilde{v}_l)}\} \\ &\quad - \sum_{l \in S1} 2 \log\{(e^{-\theta \tilde{u}_l} - 1)(e^{-\theta \tilde{v}_l} - 1) + (e^{-\theta} - 1)\} \\ &\quad + \sum_{l \in S2} \log \left\{ \frac{e^{-\theta \tilde{p}_i} - 1}{(e^{-\theta \tilde{p}_i} - 1)(e^{-\theta \tilde{v}_l} - 1) + (e^{-\theta} - 1)} \right\} \\ &\quad + \sum_{l \in S2} \log\{e^{-\theta \tilde{v}_l}\} + \sum_{l \in S3} \log\{e^{-\theta \tilde{u}_l}\} \\ &\quad + \sum_{l \in S3} \log \left\{ \frac{e^{-\theta \tilde{p}_j} - 1}{(e^{-\theta \tilde{u}_l} - 1)(e^{-\theta \tilde{p}_j} - 1) + (e^{-\theta} - 1)} \right\} \\ &\quad + \sum_{l \in S4} \log \left\{ -\theta \log \left\{ 1 + \frac{(e^{-\theta \tilde{p}_i} - 1)(e^{-\theta \tilde{p}_j} - 1)}{e^{-\theta} - 1} \right\} \right\}. \end{aligned} \quad (3.6)$$

The log-likelihood can be split into four parts, each corresponding to the contribution of observations from one of the four possible combinations (Appendix B). The notation  $\sum_{l \in S_1}$  implies summation over all the observations that fall into the first scenario,  $x_i \neq 0$  and  $x_j \neq 0$ , likewise for other summations. We define  $\tilde{u}_l = F_i(x_{li} | \tilde{\boldsymbol{\gamma}}_i)$  as the cumulative distribution function of microbe  $i$  evaluated at  $x_{li}$ , with the two-stage MLEs plugged in for the marginal parameters. The same holds for  $\tilde{v}_l$  and microbe  $j$ . Taking the derivative of Eq. 3.6 gives the two-stage score equation for  $\theta$ , which has no closed form solution (Appendix B). As a result, the two-stage MLE of  $\theta$  is found numerically using a one-dimensional optimizer in R.

### 3.3.2. Asymptotic normality

Joe (2005) obtained the asymptotic covariance matrix for the vector of two-stage estimators  $\tilde{\boldsymbol{\eta}}$  using the theory of inference functions. Specifically, by defining the inference functions

$$\mathbf{g} = (\mathbf{g}_i, \mathbf{g}_j, g_\theta)^\top, \quad (3.7)$$

where

$$\mathbf{g}_k = \frac{\partial \log f_k(\cdot | \boldsymbol{\gamma}_k)}{\partial \boldsymbol{\gamma}_k}, \quad \text{for } k \in \{i, j\} \quad (3.8)$$

and  $g_\theta = \partial \log f(\cdot | \boldsymbol{\eta}) / \partial \theta$ , it is shown that

$$\sqrt{n}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} MVN(0, \mathbf{V}), \quad \text{as } n \rightarrow \infty. \quad (3.9)$$

Where  $\mathbf{V} = (-\mathbf{D}_g^{-1}) \mathbf{M}_g (-\mathbf{D}_g^{-1})^\top$ ,  $\mathbf{M}_g = \text{Cov}(\mathbf{g}(Y | \boldsymbol{\eta})) = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]$ , and  $\mathbf{D}_g = \mathbb{E}[\partial \mathbf{g}(Y, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}^\top]$ .

Now let  $\mathcal{J} = \text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \mathbb{E}[\mathbf{g}_i \mathbf{g}_j^\top]$ ,  $\mathcal{I} = -\mathbb{E}[\partial^2 \log f / \partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_j^\top]$  and  $\mathcal{I}_{k\theta} = -\mathbb{E}[\partial^2 \log f / \partial \boldsymbol{\gamma}_k \partial \theta]$  for  $k = i, j$ . Then

$$\begin{aligned}
-\mathbf{D}_g &= \begin{bmatrix} \mathcal{J}_{ii} & 0 & 0 \\ 0 & \mathcal{J}_{jj} & 0 \\ \mathcal{I}_{\theta i} & \mathcal{I}_{\theta j} & \mathcal{I}_{\theta\theta} \end{bmatrix} & -\mathbf{D}_g^{-1} &= \begin{bmatrix} \mathcal{J}_{ii}^{-1} & 0 & 0 \\ 0 & \mathcal{J}_{jj}^{-1} & 0 \\ a_i & a_j & \mathcal{I}_{\theta\theta}^{-1} \end{bmatrix} \\
\mathbf{M}_g &= \begin{bmatrix} \mathcal{J}_{ii} & \mathcal{J}_{ij} & 0 \\ \mathcal{J}_{ji} & \mathcal{J}_{jj} & 0 \\ 0 & 0 & \mathcal{I}_{\theta\theta} \end{bmatrix}.
\end{aligned} \tag{3.10}$$

where  $a_k = -\mathcal{I}_{\theta\theta}^{-1}\mathcal{I}_{\theta k}\mathcal{J}_{kk}^{-1}$  for  $k = i, j$ .

### 3.3.3. A rescaled likelihood ratio test

In general, we are interested in determining if any two microbes  $i$  and  $j$  have a dependence structure such that  $\theta = \Theta_0$ , for some pre-specified  $\Theta_0$ . We propose a rescaled likelihood ratio test to do so. Consider the general hypothesis testing problem:

$$H_0 : \theta \in \Theta_0, \quad \text{vs.} \quad H_1 : \theta \in \Theta_1.$$

Suppose  $\ell = (\ell_i, \ell_j, \ell_\theta)^\top$  where  $\ell_i$  and  $\ell_j$  are defined above, and  $\ell_\theta = \log f(\cdot|\boldsymbol{\eta})$ . Define the two-stage likelihood ratio test statistic as:

$$\Lambda' = -2\omega[\ell(\theta_0, \tilde{\gamma}_i, \tilde{\gamma}_j) - \ell(\tilde{\theta}, \tilde{\gamma}_i, \tilde{\gamma}_j)], \tag{3.11}$$

where

$$\omega = \left( 1 + \mathcal{I}_{\theta\theta}^{-1}(\mathcal{I}_{\theta 1}\mathcal{J}_{11}^{-1}\mathcal{I}_{1\theta} + \mathcal{I}_{\theta 2}\mathcal{J}_{22}^{-1}\mathcal{I}_{2\theta} + \mathcal{I}_{\theta 1}\mathcal{J}_{11}^{-1}\mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{I}_{2\theta} + \mathcal{I}_{\theta 2}\mathcal{J}_{22}^{-1}\mathcal{J}_{21}\mathcal{J}_{11}^{-1}\mathcal{I}_{1\theta}) \right)^{-1}.$$

**Theorem 1.** *Under standard regularity conditions, we have  $\Lambda' \xrightarrow{D} \chi_1^2$ .*

The proof of Theorem 1 can be found in Appendix B. It can be shown that the above two-stage

likelihood ratio test is equivalent to the pseudolikelihood ratio test (Liang and Self, 1996). Most often, the hypothesis we are interested in testing is  $\Theta_0 = \theta_I$  where  $\theta_I$  is the value of the dependence parameter that corresponds to the independence copula. For the Frank copula, this is  $\theta_I = 0$ . Under independence, it can be shown that  $\mathcal{I}_{1\theta} = \mathcal{I}_{2\theta} = 0$ , implying that  $\tilde{\theta}$  is asymptotically efficient and the two-stage likelihood ratio statistic reduces to the regular likelihood ratio test statistic (Shih and Louis, 1995; Genest et al., 1995).

### 3.4. Simulation Studies

Simulation studies were used to assess the performance, in terms of bias and variance, of the two-stage estimation procedure, as well as the Type I error and power of the two-stage likelihood ratio test. The data was simulated using the Rosenblatt transformation, a variant of the probability integral transformation (Rosenblatt, 1952). Let  $U$  and  $V$  be defined as earlier and define a new random variable  $W$  such that,

$$W = C_{v|u}(u, v) := \frac{\partial C(u, v)}{\partial u} = Pr(V = v|U = u).$$

By the Rosenblatt transformation,  $U$  and  $W$  are independent uniform random variables and we can define the following simulation algorithm for any two microbes.

---

**Algorithm 3** Data generation from a Frank copula with mixture margins.

---

- 1: For  $l = 1, \dots, n$ :
  - 2: Simulate  $U_l \sim \text{Uniform}(0,1)$  and  $W_l \sim \text{Uniform}(0,1)$ .
  - 3: Solve for  $v_l$  by inverting the conditional copula function such that:
$$v_l = C_{w|u}^{-1}(w_l, u_l) = -\frac{1}{\theta} \log \left\{ 1 + \frac{w_l(e^{-\theta} - 1)}{w_l + e^{-\theta} u_l (1 - w_l)} \right\}$$
  - 4: Solve for  $x_{li}$  using the definition of  $U_l$ :  $x_{li} = F_i^{-1}(u_l) = \begin{cases} 0 & \text{if } u_l \leq p_{li} \\ F_{beta}^{-1}\left(\frac{u_l - p_{li}}{1 - p_{li}}\right) & \text{if } u_l > p_{li} \end{cases}$
  - 5: Repeat Step (4) for for  $x_{lj}$  and  $v_l$ .
- 

In the event that the simulation scheme above results in less than three non-zero relative abundances for either microbial taxon the procedure was repeated. This is because at least three non-zero observations are needed to be able to estimate the three taxon-specific marginal parameters. Additionally,

for any simulated data set, if the two taxa are mutually exclusive, meaning no pair of observations have non-zero relative abundance for both taxa, or if only one pair of observations has non-zero relative abundance for both taxa, the procedure was repeated. This was done because such scenarios lead to dependence parameters hitting the lower boundary of estimation and/or cause unstable variance estimates.

Simulations were performed under a variety of marginal parameter settings to understand the robustness of the estimation procedure. The dependence parameter  $\theta$  was selected from  $\{-2.5, -1, 0, 0.5, 1.5, 3\}$ . Under the marginal settings of no covariate adjustment the zero-inflation probabilities,  $(p_i, p_j)$ , were selected from  $\{(0.10, 0.25), (0.40, 0.50), (0.60, 0.75), (0.20, 0.75)\}$  and the parameters of the beta portion of the marginal distributions,  $(\mu_k, \phi_k)$ ,  $k = i, j$ , were selected from  $\{(\frac{2}{7}, 7), (\frac{5}{7}, 7), (\frac{1}{2}, 4), (\frac{1}{3}, 9), (\frac{2}{3}, 9), (\frac{1}{2}, 6)\}$ .

We also performed simulation with a single continuous covariate affecting the presence-absence probability of each microbe. Under this setting, we assumed that both  $Q_{i1}$  and  $Q_{j1}$  are drawn from a standard normal distribution, and  $p_i$  and  $p_j$  were modeled via logistic regressions,  $\text{logit}(p_k|Q_{k1}) = \rho_{k0} + \rho_{k1}Q_{k1}$ ,  $k = i, j$ , where  $Q_{k1}$  is the confounder variable,  $\rho_{k0}$  is the intercept and  $\rho_{k1}$  is the coefficient of the confounder. With corresponding vectors of true regression coefficients,  $\{(\rho_{i0}, \rho_{i1}), (\rho_{j0}, \rho_{j1})\}$ , assumed to be from one of the three following settings:  $\{(-0.5, 0.7), (-0.3, 0.4)\}$ ,  $\{(-0.1, 0.7), (0.1, 0.4)\}$ , and  $\{(0.5, 0.7), (0.8, 0.4)\}$ . In general, these models correspond to low-low, low-high, and high-high zero-inflation probabilities, respectively. The mean abundances were specified as  $\mu_i = \frac{e^{-0.7}}{1+e^{-0.7}}$  and  $\mu_j = \frac{e^{-1}}{1+e^{-1}}$  and the dispersion parameters as  $\phi_i = \phi_j = e^{1.5}$ . For all parameter settings, the sample size was set to  $n = 50$ , except under the setting with no covariates and independence ( $\theta = 0$ ) additional simulations were run for a larger sample size of 250. All simulations were repeated 500 times.

### 3.4.1. Parameter estimation

The two-stage estimator is unbiased under all dependence, zero-inflation, and marginal parameter settings (Figures 3.1 and B.1). However, under high zero-inflation, we observed some larger outliers in the estimates. This is expected since too many zeros in the data can lead to an unstable estimate

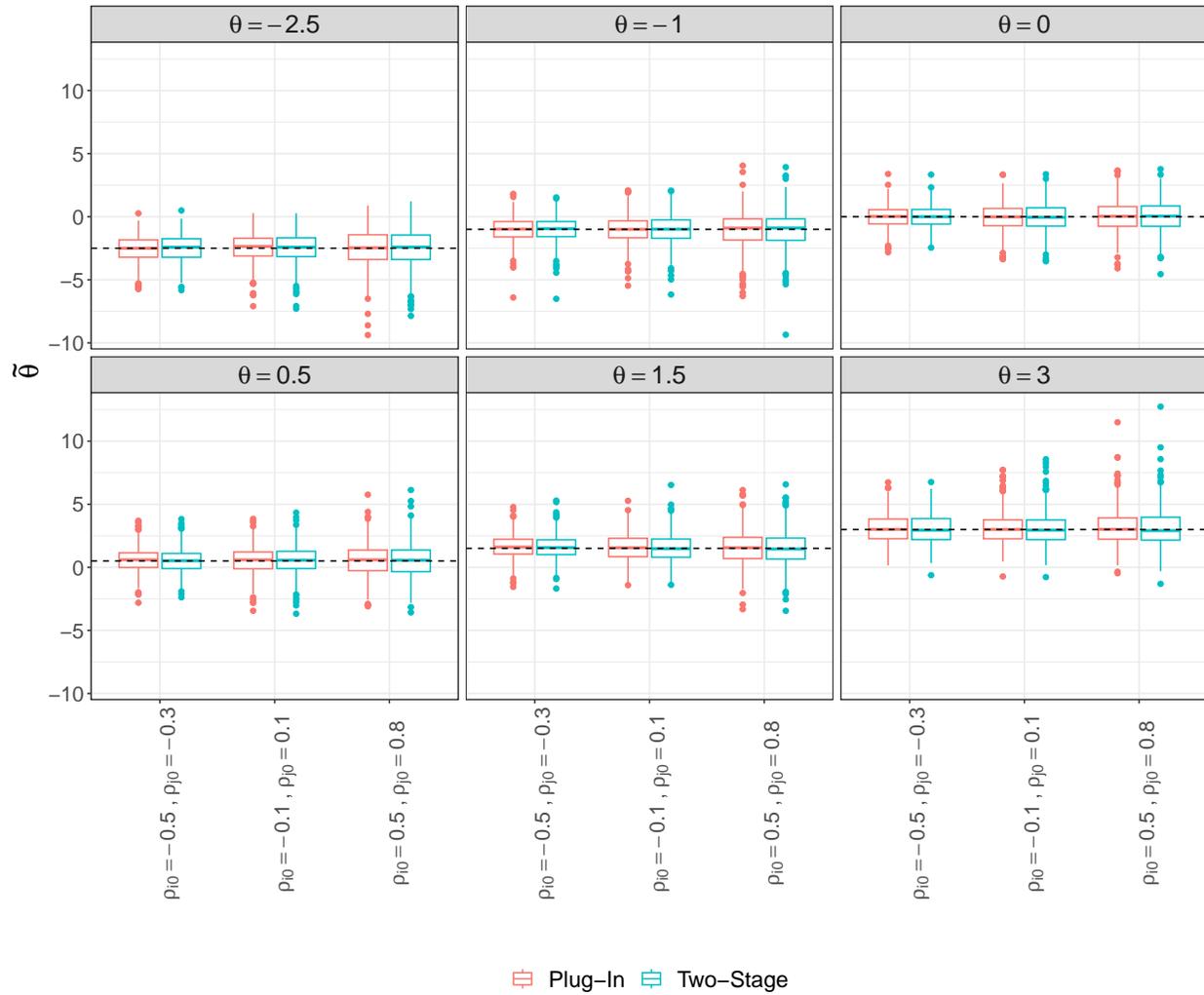


Figure 3.1: Boxplots of two-stage and plug-in estimated  $\tilde{\theta}$  across 500 simulations, where plug-in estimators use the true value of marginal parameters and perform univariate estimation of  $\theta$ . The black dashed line represents the true  $\theta$  value. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ).

of the parameters. The proposed estimator is also performed similar to a plug-in estimator that used the true value of marginal parameters and performs univariate estimation of  $\theta$ . Furthermore, the copula dependence parameter has a relationship to rank correlations such that for a given copula and  $\theta$  value the corresponding Spearman's rho ( $\rho_s$ ) and Kendall's tau ( $\tau_k$ ) can be calculated (Joe, 1997). Figure 3.2 shows for mild to strong dependence structures the typical sample estimator of Spearman's correlation is biased, even under low zero-inflation probabilities, while the copula estimator is unbiased. A similar trend holds for Kendall's tau (results not shown).

In addition to estimating  $\theta$  we also calculated its variance. The mixed partial derivatives necessary to calculate covariance matrix,  $\mathbf{V}$ , are analytically difficult to compute, therefore we replace it with a consistent estimator, such as the jackknife estimator:

$$n^{-1}\tilde{\mathbf{V}} = \sum_{l=1}^n (\tilde{\boldsymbol{\eta}}^{(l)} - \tilde{\boldsymbol{\eta}})^\top (\tilde{\boldsymbol{\eta}}^{(l)} - \tilde{\boldsymbol{\eta}}).$$

The variance of  $\theta$  is the [7, 7] entry of  $n^{-1}\tilde{\mathbf{V}}$ , denoted as  $\hat{\sigma}_\theta^2$ , and  $\tilde{\boldsymbol{\eta}}^{(l)}$  is a vector of two-stage maximum-likelihood estimates calculated with the  $l^{th}$  observation removed. In general, the variance increased as zero-inflation increased, regardless of dependence or marginal parameter values (Figure B.2). Specifically, without adjusting for covariates, under high zero-inflation of both microbes and moderate-to-strong positive dependence, there was an increase in large outlier estimates (Figure B.3). These results show that the mean of the analytical variance is typically larger than the empirical (sample) variance of  $\tilde{\theta}$  across all 500 simulations (Figure B.4). Though the latter almost always fell within the standard error of the former. The difference between the two increases with zero-inflation. This indicates that the jackknife estimator is conservative (upwardly biased) and may lead to a two-stage likelihood ratio test that is conservative as well. As expected, as the sample size increases the variance decreases across the board, though the same trends are seen (results not shown).

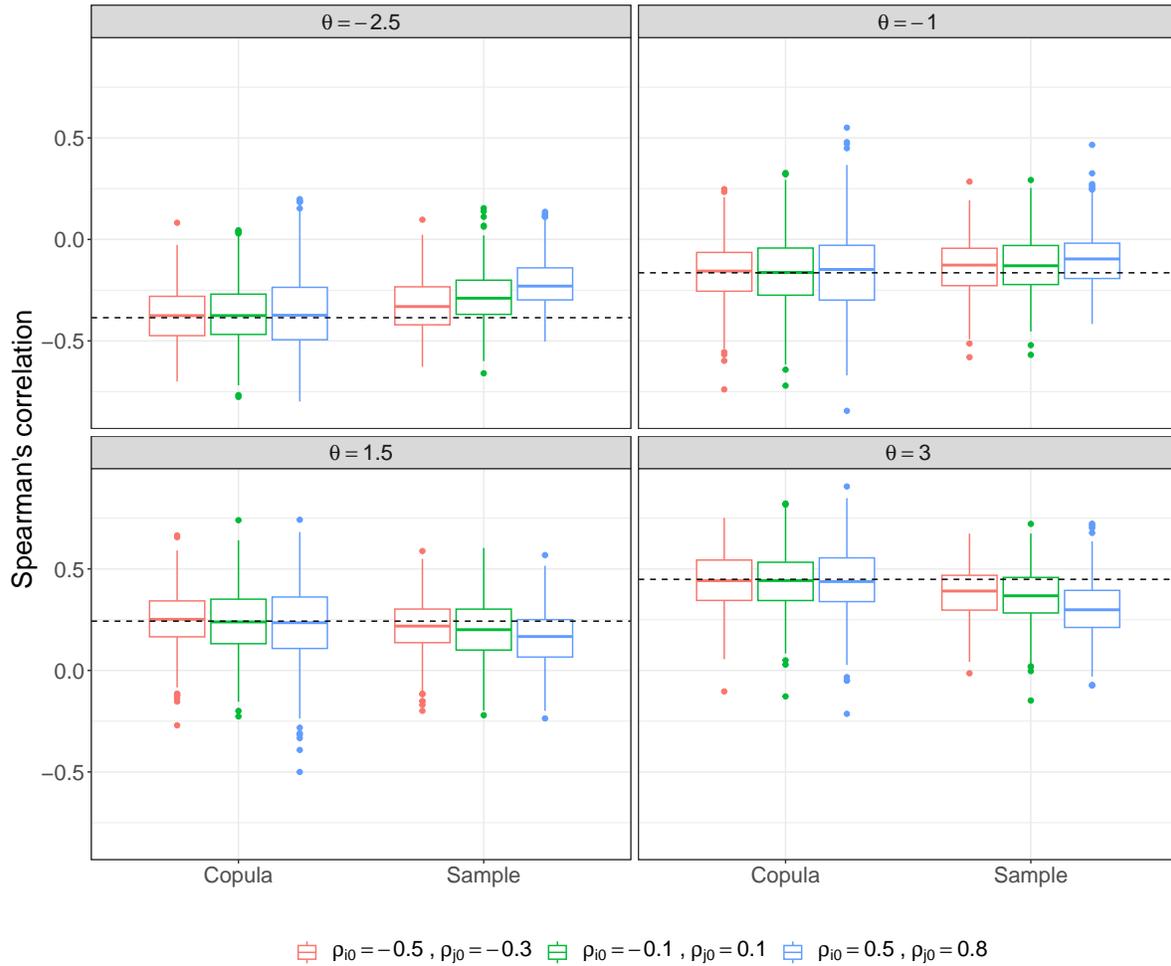


Figure 3.2: Boxplots of estimated Spearman's correlation, using copula and sample estimators, across 500 simulations. The black dashed line represents the true value. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ).

### 3.4.2. Type I error and power

We are also interested in assessing the Type I error and power of the two-stage likelihood ratio test. Specifically, we would like to test the null hypothesis that two microbes are independent (i.e.  $H_0 : \theta = 0$  for the Frank copula) versus the general two-sided alternative hypothesis that the two microbes are not independent (i.e.,  $H_1 : \theta \neq 0$ ). For the setting with covariate adjustment, our proposed likelihood ratio test uniformly outperforms standard sample correlation tests for independence using Pearson’s correlation, as well as Spearman’s and Kendall’s tau rank correlation (Figure 3.3). Under low to moderate zero-inflation, as the absolute value of the true  $\theta$  moves away from zero, in either direction, the power of the test increases symmetrically. This does not hold under dual high zero-inflation where the power to detect a true positive dependence structure increases much more rapidly than that of a true negative dependence structure. This trend does not hold in the setting without covariates (results not shown), under which the four tests performed comparably. This is likely due to the unique mapping between  $\theta$  and Spearman’s and Kendall’s tau rank-based correlations in such settings. Though, there is a slight improvement in our proposed method under dual-high zero-inflation, which corresponds to the setting where sample estimators of rank correlations are biased towards the null value of zero.

### 3.4.3. Model robustness and comparison

We further investigated the robustness of the proposed method in terms of its ability to recover dependency structures under copula model misspecification. First, we simulated data from an underlying bivariate Gaussian copula, instead of the Frank copula that we use in our method. Model parameters were specified in the same way described previously, assuming covariate adjustment. Second, we simulated data under a multivariate Gaussian copula with zero-beta mixture margins. We set the number of margins to 75 and the simulated data was normalized to have unit sum within samples, emulating microbial relative abundance data. More details on parameter specification can be found in Appendix B.

We find that even under model misspecification, the proposed method outperforms sample estimators of rank correlations (Figure B.5), resulting in less biased estimates of the model parameters

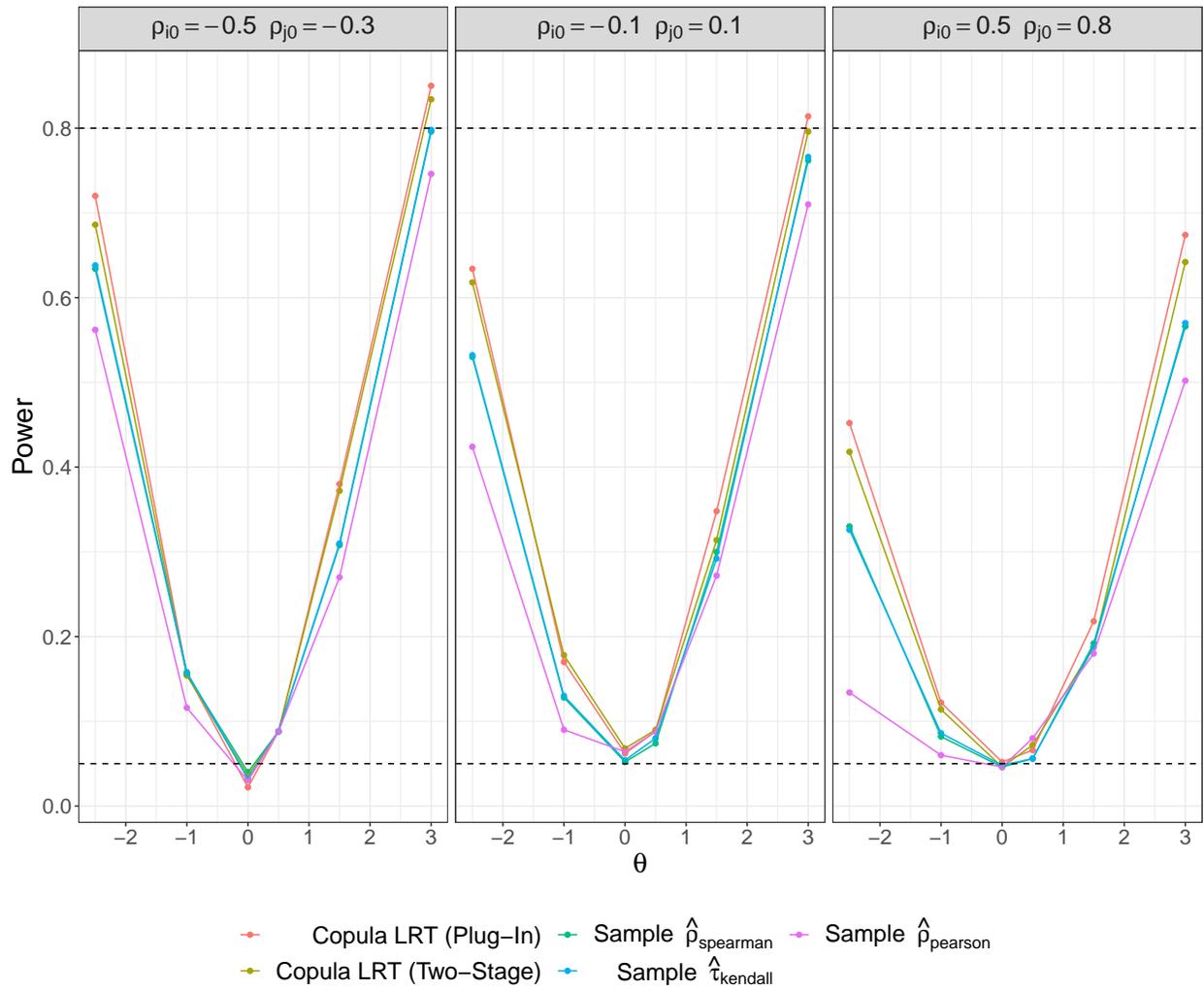


Figure 3.3: Power curves of the two-stage likelihood ratio test for independence compared to sample correlation tests. Black dashed lines at 0.05 (Type I error) and 0.8. Power was calculated under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ) with covariate adjustment.

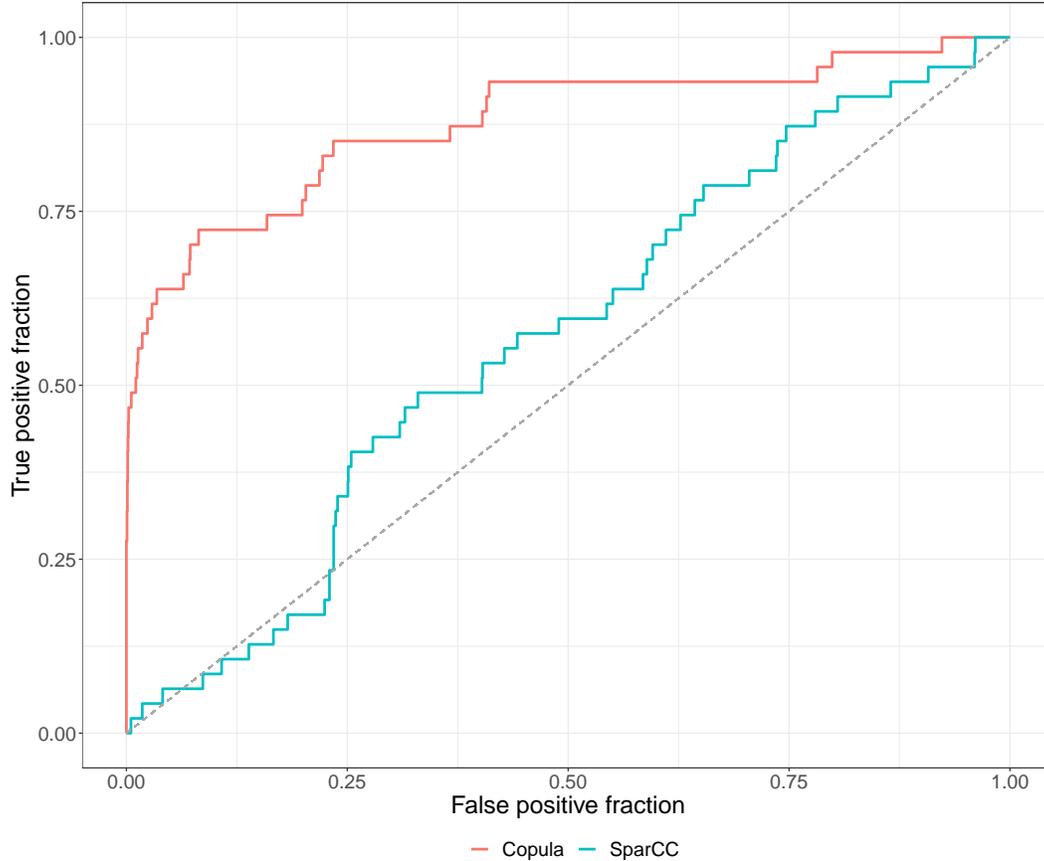


Figure 3.4: ROC curves based the two-stage estimator of the copula dependence parameter and SparCC under multivariate Gaussian copula simulation. Cutoffs were selected based on two-stage likelihood ratio test p-values for the copula model and the absolute value of the estimated correlation for SparCC, as suggested in the original manuscript.

and the corresponding Spearman’s correlations. Using multivariate simulation we investigated the proposed method’s ability to correctly identify significant pairs, across different p-value cutoffs. We compared our method to SparCC, as it also focuses on identifying marginal associations using microbial relative abundance data (Kurtz et al., 2015). Figure 3.4 shows the copula estimator performed well in recovering the pairs with a true dependency ( $AUC = 0.88$ ), outperforming SparCC which performed only slightly better than random chance ( $AUC = 0.57$ ). Though, it is important to note that SparCC’s parameter of interest is different than the proposed method’s as it aims to make inference on covariation of the unobserved absolute abundances of the taxa.

### 3.5. Analysis of a Microbial Covariation Network in the Healthy Human Gut

#### 3.5.1. Identification of pairwise microbial covariations

We used data from the self-selected and open-platform cohort of the American Gut Project (AGP) to test our proposed method (McDonald et al., 2018). The AGP cohort is made up of participants who opted into the study by giving informed consent, as well as paying a fee to cover the cost of sample processing and sequencing. The majority of the samples are from individuals living in the United States, though some samples are from individuals living in the United Kingdom or Australia. The self-reported metadata and 16S rRNA gene sequencing data are accessible from European Bioinformatics Institute under accession number ERP012803.

The data consisted of fecal microbiome samples from 3679 citizen-scientists and 971 unique genera. We filtered the sequencing data such that any reads that were unassigned at the genus level classification were removed. Any genera with a prevalence of less than 20% across all subjects were removed as well. This left a total of 68 genera for downstream analyses. Furthermore, any samples that had total number of reads of zero after the aforementioned filtering were removed. Since the data also included self-reported metadata, we adjusted for known confounders of the gut microbiome in the marginal regression models. In particular, we adjusted for age ( $44.6 \pm 17.4$ ), bmi ( $23.9 \pm 5.26$ ) and antibiotic use (69% not in the last year, 14% in the last year, 13% in the last six months, 2% in the last month, and 2% in the last week). Due to the low rate of missing data for each, less than 5% for age and antibiotic use and about 10% for BMI, we performed a complete case analysis. We further restricted our sample of interest to “healthy” individuals, defined as those who reported not having inflammatory bowel disease or diabetes, as both are known to be associated with dysbiosis. This left 2754 samples remaining.

From these 68 genera we formed 2278 unique pairs. For each of these pairs, we performed two-stage maximum likelihood estimation of the parameters and a likelihood ratio test for independence. Due to the large number of pairwise tests, we adjusted for multiple comparisons by controlling the false discovery rate at 1% level. In particular, since the test statistics are not independent from one

another we used the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001). After FDR control we identified 1314 pairs of taxa with a significant dependence among healthy subjects.

We compared the results from our method to those from Pearson’s correlation, which detected 276 significant pairs. The two methods have 233 pairs in common, our proposed method identified 1081 pairs that Pearson’s correlation did not, and Pearson’s correlation identified 43 pairs not detected by our method. The pairs detected by Pearson’s method, but not by the proposed method have a Pearson’s correlation of between -0.16 and 0.12. The copula estimate of Spearman’s correlation for these pairs are between -0.08 and 0.09, only three of the pairs detected by our method fall in this range. We also compared to SparCC, which detected 86 pairs, all of which were also identified by the copula LRT. Hierarchical clustering of the SparCC adjacency matrix, weighted by the estimated correlation, shows the method misses many of biologically relevant pairs found by the proposed method. The two clusters identified share some similarities to the clusters in Figure 3.5, but largely miss pairs between taxa from the same phylum. Lastly, we did not compare to sample rank correlation methods, as simulations showed that in the presence of excessive zeros their estimation procedure is biased.

### 3.5.2. Properties of microbial covariation network in healthy human gut

We used the results from the likelihood ratio test for independence to construct a microbial covariation network and adjacency matrix. More specifically, each microbe is a node in the network and two nodes are said to have an edge, or connection, if the result from the microbe pair’s LR test for independence was significant (FDR p-value < 0.01). Otherwise, the two nodes are said to be unconnected. A heatmap of the weighted adjacency matrix, where the weight is the estimated  $\theta$  value, shows the covariation relationships between all microbial pairs (Figure 3.5). The heatmap was clustered using complete agglomerative hierarchical clustering. Figure 3.5 shows that the network consists mostly of pairs with positive dependence, especially within clusters, with some negative dependencies between a small set of taxa, mostly between clusters. Furthermore, the microbes (nodes) of the network form three distinct clusters, identified by cutting the dendrogram from hierarchical clustering. The most common phylum in each cluster was *Firmicutes*, *Proteobacteria*, and

*Bacteroidetes*. This implies that the clusters have a biological interpretation with taxa of the same phylum tending to be members of the same cluster.

To summarize the resulting network, we calculated the average of node-specific network summary statistics. The network has an average degree of 0.577 (sd=0.146), average closeness of 0.710 (sd=0.072), average betweenness of 0.006 (sd=0.004). The high average degree of the nodes implies the network is dense with many connections. This is further implied by the network's edge density of 0.58. Meanwhile, the high eigenvalue centrality of 0.704 (sd=0.198) implies that well connected nodes are likely to be connected with each other. The network also has a diameter of 2 and a mean distance of 1.42.

We simulated 1000 random graphs from the Erdős-Rényi model with the same number of links as the observed AGP network to compare global network measures from these graphs to that of the AGP network. Both the average cluster coefficient (0.695) and modularity (0.137) of the AGP network were significantly different from those of the random graphs ( $p < 0.001$ ). Thus implying that the network structure and clusters are not formed due to clustering of random noise in the data. Additionally, we compared the cumulative degree distribution of the AGP network to that of the 1000 random graphs. We observed that the distribution of the random graphs begins around 35 degrees and increases steeply until it levels off at 50 degrees. In contrast, the distribution of the AGP network begins early around 20 degrees and rises slowly until a maximum of approximately 60 degrees.

### 3.5.3. Consistency analysis

To assess the robustness and consistency of the identified microbial pairs to slight changes in the observed data we took 50 bootstrap samples of the relative abundance data, then repeated the estimation and testing analyses, including FDR control at the 0.01 level. If the identified pairs are truly associated with one another we should see high consistency, or overlap, in the identified pairs between the original data and bootstrap samples. The average number of significant dependent pairs of taxa across all bootstrap samples, rounded to the nearest integer, is 1335. The minimum number of identified pairs is 1274 and the maximum is 1393. The average overlap and Dice coeffi-

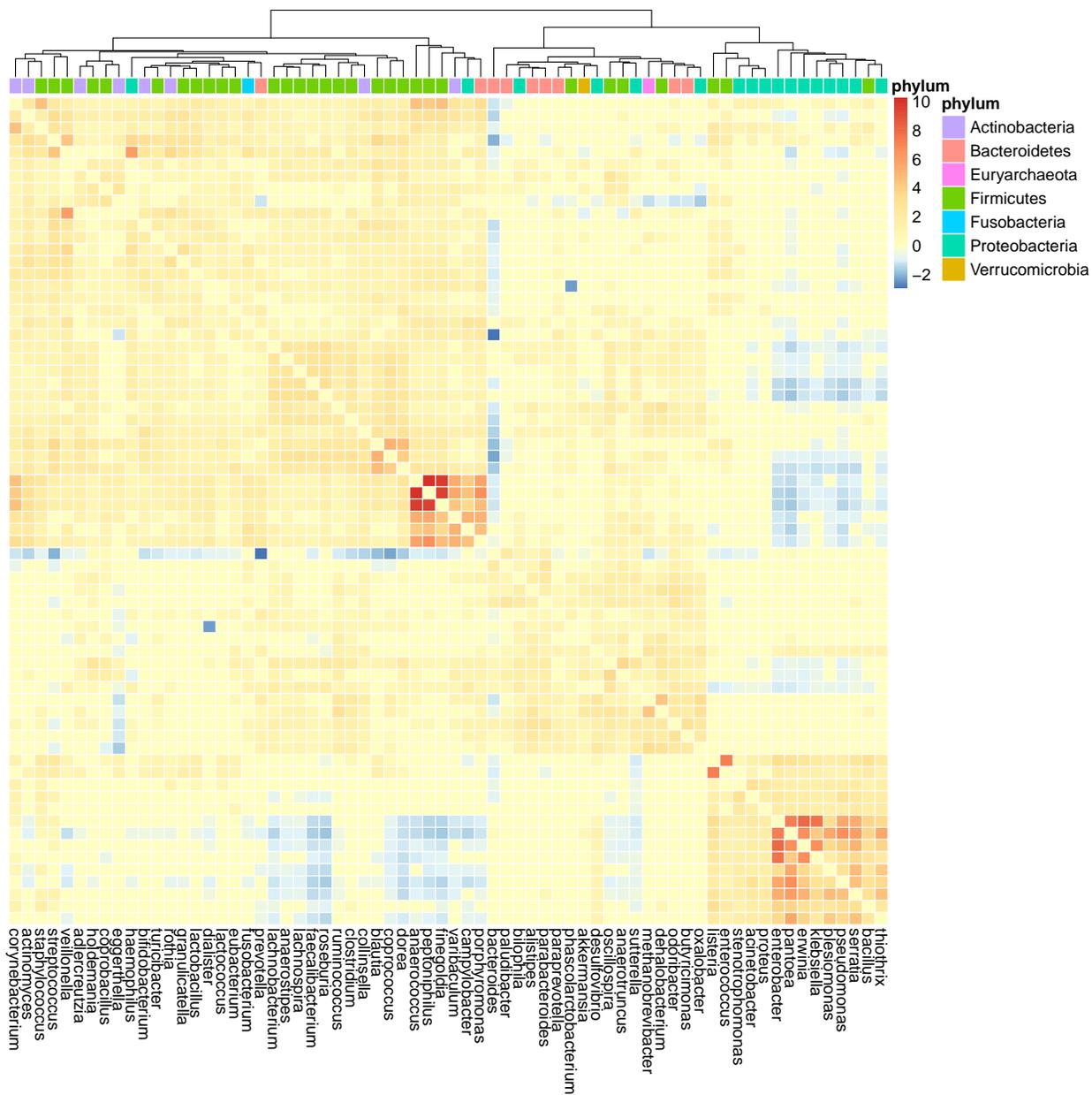


Figure 3.5: Heatmap of the AGP adjacency matrix with three clusters identified by complete agglomerative hierarchical clustering. Red indicates positive covariation and blue negative covariation.

cients between the pairs identified in the original data and those of each bootstrap sample is 0.940 (sd=0.010) and 0.930 (sd=0.006), respectively. This indicates that the identified significant pairs are robust to small changes in the observed data. Furthermore, of the 1314 microbial pairs identified from original data, 875 of these pairs were also identified in all 50 bootstrap samples and 1071 pairs were identified in over 90% of the bootstrap samples. Only 14 were identified in less than half of the bootstrap samples.

### 3.6. Discussion

In this chapter we described a bivariate copula based density for microbial relative abundance data using zero-inflated beta margins. As such, this allows for a two-stage maximum likelihood estimation procedure and corresponding two-stage likelihood ratio test for the copula dependence parameter. Such tests can be used to identify the covarying pairs of bacteria and the corresponding covariation network. Using zero-beta mixture margins provides a flexible way to capture the characteristic excess of zeros in microbiome data, allows for covariate adjustment via the margins, and uncertainty quantification of the estimator of the dependence parameter.

The low bias and high efficiency of the proposed two-stage estimator of the dependence parameter under unknown margins is a valid, and less computationally intensive, alternative to full maximum likelihood estimation. We benchmarked the runtime of our estimation and testing procedure for the copula dependence parameter. Our two-stage algorithm ran in just under 3 minutes for a sample size of 100 and in about 7 minutes for sample size of 500 on a MacBook Pro with a 2.9 GHz 6-Core Intel Core i9 processor using 6 cores. We extend current work on copula models with mixed margins, as well as work on copula two-stage estimation with our proposed two-stage likelihood ratio test (Gunawan et al., 2020; Shih and Louis, 1995; Joe, 2005). Simulation studies show under the independence hypothesis the test controls Type I error and is more powerful than tests based on sample correlation measures.

While this chapter focuses on the Frank copula, the methods discussed are quite general and hold for any Archimedean copula function. For example, the Gaussian,  $t$ -, and Clayton copulas can model positive and negative dependence as well but all assume tail dependence, which the Frank

copula does not. Additional extensions of this work include modifications to handle longitudinal data to understand the changes in microbial dynamics. A particular extension to estimate conserved covariation networks is discussed in Chapter 4.

## CHAPTER 4

# MIXTURE MARGIN RANDOM-EFFECTS COPULA MODELS FOR INFERRING TEMPORALLY CONSERVED MICROBIAL COVARIATION NETWORKS FROM LONGITUDINAL DATA

### 4.1. Introduction

A microbiome, and its set of interactions, form a complex and dynamic ecosystem (Gerber, 2014; Li, 2015). Such systems are emergent; they are characterized by properties that arise from the interactions between parts, but not by any individual component (Lidicker, 1979). It is useful to measure these emergent properties over time since it is likely that their short- and long-term effects differ. The rise in longitudinal microbial sequencing studies provides such an opportunity, as they allow for the examination of the natural variability of a microbiome. This is in contrast to cross-sectional studies, which have been successful in microbial diversity and differential abundance analyses (Xu and Knight, 2015; He et al., 2015; Paulson et al., 2013; White et al., 2009). Though, they are limited in that they only provide a snapshot of microbial dynamics at a single point in time. Longitudinal studies allow for full reconstruction of dynamics through time.

Longitudinal microbiome data allows us to identify temporally conserved microbial covariation networks while allowing for microbial compositions to change over time. To illustrate such temporally conserved microbial covariations, Figure 4.1 shows two pairs of microbes from the gut microbiome of some children in the Broad Institute’s DIABIMMUNE pediatric cohort (Yassour et al., 2016). While their relative abundances may change over time, the overall dependency structure and pattern of the plots remains the same. These consistent patterns are the conserved covariations we would like to capture. The DIABIMMUNE antibiotics cohort consists of 39 children from Finland. The gut microbiome of each child is densely sampled with an average of 28 samples per child collected monthly over the first 36 months after birth. Twenty children received between nine and fifteen antibiotic treatments and the other 19 were never exposed to antibiotics, thus allowing for comparison between the two groups in terms of their conserved microbial covariation networks.

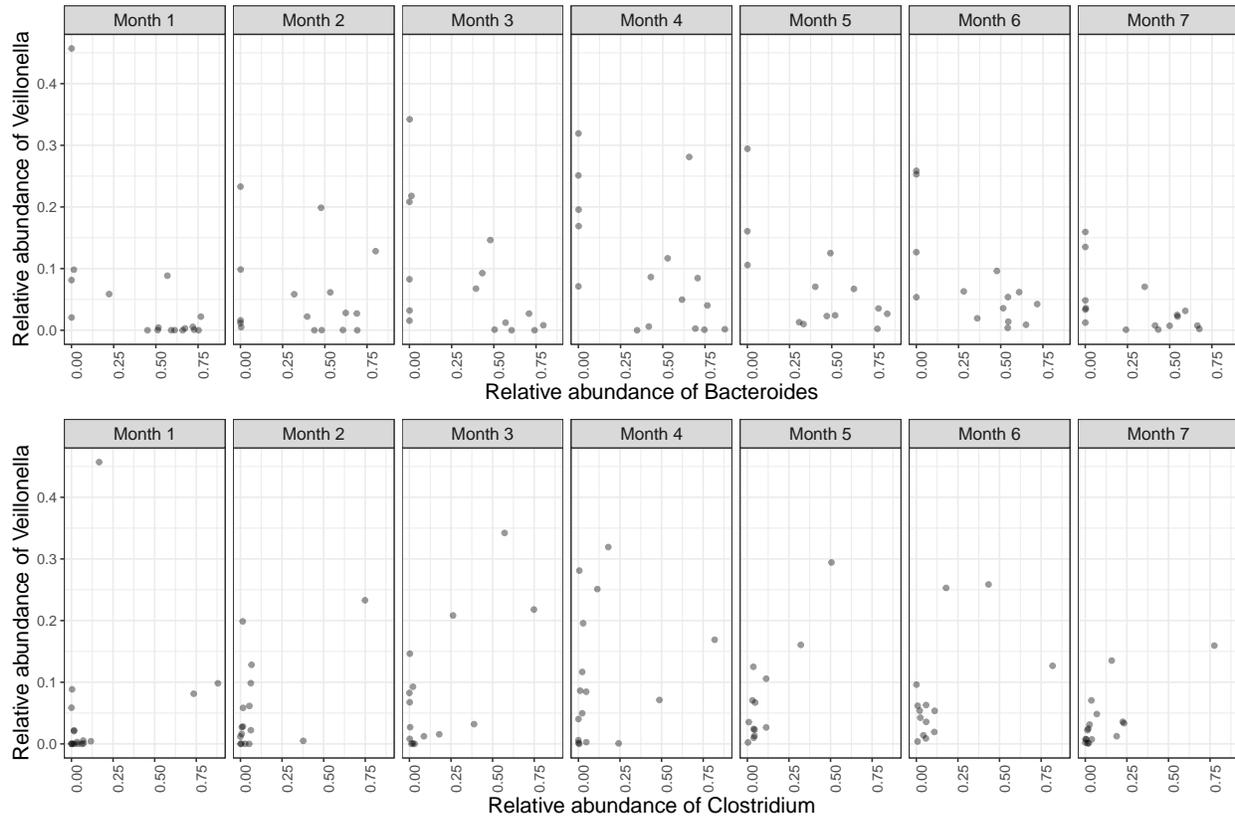


Figure 4.1: Scatterplots of the proportion relative abundance of two pairs of bacteria (top: *Veillonella* and *Bacteroides*, bottom: *Veillonella* and *Clostridium*) over the first seven months of life for children unexposed to antibiotics in the DIABIMMUNE study (Yassour et al., 2016). The general covariation structure is consistent over the seven months. A negative covariation for *Veillonella* and *Bacteroides* and positive covariation for *Veillonella* and *Clostridium*, illustrating a conserved dependence structure between the taxa.

This chapter presents a model based method for identifying temporally conserved microbial covariation networks. We are interested in quantifying the conserved covariation structure as a measure of biological robustness. Defined as a property that allows a system to maintain function in the presence of perturbations, robustness is a key property of dynamic and evolvable systems such as the microbiome (Kitano, 2004, 2007; Félix and Barkoulas, 2015). There has been interest in studying robustness in a systems biology context, particularly in cellular and molecular systems (Stelling et al., 2004). Since biological robustness is a trait-specific property, there are many ways to measure microbial robustness (Besten et al., 2010; Olsson et al., 2022). Dynamics that are conserved through time are one way to measure robustness, as they may provide information regarding the organization, structure and function of the microbiome. Raman et al. (2019) proposed studying microbial covariations that are stable through time via conserved covariance, defined as the average of the binary (thresholded) covariance matrices across all observed time points. In particular, the method identifies a group of consistently covarying taxa using the conserved covariance matrix and shows that this group can be used to distinguish children based up on malnutrition status. However, the method is limited by the fact that microbial relative abundance data are often very sparse with many zeros. Standard measures of pairwise association, such as Pearson’s correlation and pairwise covariance, can lose power or lead to biased estimates of associations in this setting.

We focus on the problem of estimating the conserved pairwise association between two bacterial taxon based on zero-enriched relative abundance data, which are the key features of microbiome data. Specifically, we define a general conserved dependence parameter between microbial pairs using generative copula models with mixture margins for the normalized relative abundance data. Copula models are advantageous because they separate the modeling of the dependence structure from that of the univariate margins. Mixture margin copula models have been shown to fit sparse microbiome proportion data well (Deek and Li, 2021). However, application of copula models to longitudinal data, outside of vine copulas, has been limited. We propose a random-effects model for the copula dependence parameter that captures conserved microbial covariations while allowing for time-specific parameters in the copula model. To this end, we assume the copula dependence parameters across different time points follow a Gaussian distribution, whose mean parameter can

be used to quantify the conserved microbial covariations and to build conserved covariation networks. The distribution’s variance parameter controls the variability of the time-specific dependence parameters around the temporally conserved mean.

To estimate the model parameters, we propose a Monte Carlo EM algorithm where Metropolis-Hastings sampling is used in the E-step. We also develop a Monte Carlo likelihood ratio test for hypothesis testing of the conserved dependence parameter. Our simulations show that the MCEM algorithm is efficient and provides a good estimate the conserved dependence parameter in the random-effects copula models. We also show using simulations that the Monte Carlo likelihood ratio test has the correct Type I error and outperforms naive correlation methods. Finally, we present a detailed analysis of conserved microbial covariation networks in the infant gut microbiome and evaluate the effects of antibiotics on microbial covariation network robustness, stability and centrality.

## 4.2. Mixture margin random-effects copula models and Monte Carlo inference

### 4.2.1. Mixture margin random-effects copula models for longitudinal data

Consider a microbial sample taken from an individual at time  $t$ , that can be summarized by a vector of taxon counts:  $(y_1^{(t)}, \dots, y_m^{(t)}) \in \mathbb{R}_+^m$ , where  $m$  is the number of microbial features. Due to sequencing constraints, these counts only contain information on the relative abundance of each taxon, thus making it difficult to compare them across samples. For this reason, the counts are typically normalized by the total number of sequencing reads of the sample. That is, the normalized relative abundances of the  $m$ -microbes observed is defined as  $x_m^{(t)} = \frac{y_m^{(t)}}{\sum_m y_m^{(t)}}$  and denoted by the vector:  $(x_1^{(t)}, \dots, x_m^{(t)}) \in [0, 1]^m$ .

Our focus is on measuring and testing covariation among pairs of the microbes. Specifically, for any two of the microbes,  $i$  and  $j$ , let the joint cumulative distribution function of their relative abundances measured at time  $t$  have the general copula form:

$$F(x_i^{(t)}, x_j^{(t)} | \gamma_i^{(t)}, \gamma_j^{(t)}, \theta_{ij}^{(t)}) = C(F_i(x_i^{(t)}; \gamma_i^{(t)}), F_j(x_j^{(t)}; \gamma_j^{(t)}) | \theta_{ij}^{(t)}) = C(u^{(t)}, v^{(t)} | \theta_{ij}^{(t)}),$$

where  $F_k(\cdot; \gamma_k^{(t)})$  is the  $k$ -th univariate margin ( $k = i, j$ ) with parameters  $\gamma_k^{(t)}$  and  $C(\cdot|\theta^{(t)})$  is a family of copulas with dependence parameter  $\theta^{(t)}$ , at time  $t$ . Henceforth, we omit the subscripts in the dependence parameters to ease notation, implying that we are referring to the modeling of a given pair  $(i, j)$  of microbes, unless otherwise noted. The univariate margins for microbial relative abundance data can be described by the zero-inflated beta distribution,

$$f_k(x_k^{(t)}) = p_k^{(t)} I_k^{(t)} + (1 - p_k^{(t)}) f_{beta}(x_k^{(t)} | \mu_k^{(t)}, \phi_k^{(t)}) (1 - I_k^{(t)}),$$

which has proven to be a powerful and robust parametric distribution in the modeling of such data (Peng et al., 2015; Chen and Li, 2016). We define  $p_k^{(t)} = \Pr(x_k^{(t)} = 0)$  as the zero-inflation probability of microbe  $k$  at time  $t$ ,  $I_k^{(t)} = I(x_k^{(t)} = 0)$ , and

$$f_{beta}(x_k^{(t)} | \mu_k^{(t)}, \phi_k^{(t)}) = \frac{\Gamma(\phi_k^{(t)})}{\Gamma(\mu_k^{(t)} \phi_k^{(t)}) \Gamma((1 - \mu_k^{(t)}) \phi_k^{(t)})} x_k^{\mu_k \phi_k - 1} (1 - x_k^{(t)})^{(1 - \mu_k^{(t)}) \phi_k^{(t)} - 1},$$

the density function of a beta random variable indexed by mean parameter  $\mu_k^{(t)}$  and dispersion parameter  $\phi_k^{(t)}$ .

The advantage of using a copula model is that the univariate marginal distribution allows for adjustment of possible covariates  $\mathbf{z}_k^{(t)}$  in modeling  $F_k(x_k^{(t)}; \gamma_k^{(t)})$ . Here, the marginal parameters  $\gamma_k^{(t)}$  include the regression coefficients from logistic or log-linear regression models for  $p_k^{(t)}$ ,  $\mu_k^{(t)}$ , and  $\phi_k^{(t)}$ . In longitudinal microbiome studies,  $\mathbf{z}_k^{(t)}$  can include the relative abundance of the  $k^{th}$  microbe observed at a set of prior time points  $\{x_k^{(t-1)}, \dots, x_k^{(t-p)}\}$ , which leads to auto-regressive marginal models.

To define our proposed mixture margin random-effects copula models, we assume that the dependence parameter at time  $t$  follows a Gaussian distribution with mean  $\theta$  that captures the time-invariant dependence between any two microbes,

$$\theta^{(t)} = \theta + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2),$$

where  $\sigma^2$  is the variance parameter that controls the variability of the time-specific dependence parameters ( $\theta^{(t)}$ ) around the mean ( $\theta$ ). Specifically, for any two microbes, this model allows for time-specific dependence parameters that vary around  $\theta$ , the conserved dependence parameter, which is used to measure the microbe-pair's conserved covariation. As the magnitude of  $\sigma^2$  increases, the variability of the time-specific dependence parameters around  $\theta$  increases.

Based on the value of  $\theta$  and  $\sigma$  we define four types of temporal covariation. The first is when both  $\theta$  and  $\sigma$  are equal to zero, which is defined as no covariation. This is because when both parameters are zero every time-specific copula dependence parameter is also equal to zero, the independence value for the Frank copula. The second is when  $\theta$  is zero but  $\sigma$  is nonzero. We define this as no conserved covariation because at any given time point the taxa can be associated with one another ( $\theta^{(t)} \neq 0$ ), but the covariation may be positive or negative and varies randomly around the independence value. In the last two settings,  $\theta$  is nonzero and thus there is a conserved covariation structure. When  $\sigma$  is zero we have strictly conserved covariation, as there is no variability around the conserved value, the dependence structure is the same at every time point. Whereas when  $\sigma$  is nonzero we have a more relaxed definition of conserved covariation that allows for some variability in the time-specific dependence parameters around the conserved value.

We are interested in estimating the parameter  $\theta$  and in performing hypothesis testing of  $H_0 : \theta = \theta_0$ , where  $\theta_0$  is some pre-specified, null value. In real microbiome applications, we test  $H_0 : \theta_{ij} = \theta_0$  for each pair of microbes  $i$  and  $j$  and adjust for multiple comparisons by controlling the false discovery rate at a given level. We are most interested in the case where  $\theta_0 = 0$ , when Frank copula is used, which corresponds to independence between two microbes. We can then construct a conserved covariation network of all the microbes by creating links among those pairs with  $\theta_{ij} \neq 0$ .

#### 4.2.2. Joint density and the likelihood function

Given  $\theta^{(t)}$ , let the bivariate joint density function be given by  $f(x_i^{(t)}, x_j^{(t)} | \gamma_i^{(t)}, \gamma_j^{(t)}, \theta^{(t)})$ , which can be found via appropriate differentiation of the copula function  $C$ . When the margins are mixtures of absolutely continuous and discrete random variables, as is the case with the zero-inflated beta density, special care must be taken in the differentiation of the copula distribution function. A

general framework for finding the density in such situations has been defined, as well as the density for the special case of bivariate zero-inflated beta margins (Gunawan et al., 2020; Deek and Li, 2021). For each time point, the marginal distribution function is given by

$$f(\mathbf{x}^{(t)}|\boldsymbol{\gamma}^{(t)}, \theta, \sigma^2) = \int f(\theta^{(t)}|\theta, \sigma^2) \prod_{l=1}^n f(\mathbf{x}_l^{(t)}|\boldsymbol{\gamma}^{(t)}, \theta^{(t)})d\theta^{(t)}, \quad (4.1)$$

where  $\mathbf{x}_l^{(t)} = (x_{il}^{(t)}, x_{jl}^{(t)})$  and  $\boldsymbol{\gamma}^{(t)} = (\gamma_i^{(t)}, \gamma_j^{(t)})$ . Additionally, assume the data from each of the  $t$  time points ( $t = 1, \dots, T$ ) are conditionally independent and therefore the full data likelihood can be found by taking their product,

$$f(\mathbf{X}|\boldsymbol{\Gamma}, \theta, \sigma^2) = \prod_{t=1}^T \int f(\theta^{(t)}|\theta, \sigma^2) \prod_{l=1}^n f(\mathbf{x}_l^{(t)}|\boldsymbol{\gamma}^{(t)}, \theta^{(t)})d\theta^{(t)}, \quad (4.2)$$

where  $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)})$  and  $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}^{(1)}, \dots, \boldsymbol{\gamma}^{(T)})$ . It is important to note that the integration required in Eq. 4.2 is intractable. Thus, an approximate method for parameter estimation must be used.

#### 4.2.3. A two-stage estimation procedure

Under our proposed mixture margin random-effects model, the parameters of inferential interest are the true time-invariant dependence parameter  $\theta$  and the variance,  $\sigma^2$ , of the random-effect whereas the parameters of the time-dependent marginals,  $\boldsymbol{\Gamma}$ , can be regarded as nuisance parameters. For ease of notation we define the full log-likelihood as

$$\ell(\theta, \sigma^2) = \ell(\theta, \sigma^2, \boldsymbol{\Gamma}) = \log f(\mathbf{X}|\boldsymbol{\Gamma}, \theta, \sigma^2). \quad (4.3)$$

Consequently, we adopt a two-stage, or inference-for-margins, approach that is commonly used for inference involving copula models (Shih and Louis, 1995; Joe and Xu, 1996). The general two-stage estimation scheme can be described as follows:

1. Separately maximize each of the  $2T$  univariate log-likelihoods,

$$\ell_k(\boldsymbol{\gamma}_k^{(t)}) = \sum_{l=1}^n \log f_k(x_{lk}^{(t)} | \boldsymbol{\gamma}_k^{(t)}) \quad k = i, j \text{ and } t = 1, \dots, T,$$

to get estimates of their parameters  $\tilde{\boldsymbol{\gamma}}_k^{(t)}$ .

2. Maximize the function  $\ell(\theta, \sigma^2, \tilde{\boldsymbol{\Gamma}})$  over  $\theta$  and  $\sigma^2$  to get estimates  $\tilde{\theta}$  and  $\tilde{\sigma}^2$ .

Two-stage estimation reduces the computational complexity of the estimation procedure by replacing a single, joint maximization problem of many parameters with several smaller maximizations. However, the second step of the algorithm cannot be solved directly due to the intractable integral in the likelihood function. Accordingly, we propose a Monte Carlo Expectation-Maximization (EM) algorithm to do so.

#### 4.2.4. A Monte Carlo Expectation-Maximization algorithm

The Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) is particularly well-suited to solve the maximization of the marginal log-pseudolikelihood,  $\tilde{\ell}(\theta, \sigma^2) = \ell(\theta, \sigma^2, \tilde{\boldsymbol{\Gamma}})$ . Under this framework,  $\mathbf{X}$  are the observed data,  $\theta$  and  $\sigma^2$  are the unknown but fixed parameters of interest, and  $\boldsymbol{\Theta}^{(T)} = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)})$  are unobserved latent variables or missing data. We define  $(\mathbf{X}, \boldsymbol{\Theta}^{(T)})$  as the complete data and  $\tilde{\ell}(\theta, \sigma^2, \tilde{\boldsymbol{\Gamma}} | \mathbf{X}, \boldsymbol{\Theta}^{(T)}) = f(\mathbf{X}, \boldsymbol{\Theta}^{(T)} | \tilde{\boldsymbol{\Gamma}}, \theta, \sigma^2)$  as the complete log-pseudolikelihood function given by

$$\tilde{\ell}(\theta, \sigma^2, \tilde{\boldsymbol{\Gamma}} | \mathbf{X}, \boldsymbol{\Theta}^{(T)}) = \sum_{t=1}^T \log f(\theta^{(t)} | \theta, \sigma^2) + \sum_{t=1}^T \sum_{l=1}^n \log f(\mathbf{x}_l^{(t)} | \boldsymbol{\gamma}^{(t)}, \theta^{(t)}). \quad (4.4)$$

The assumption of the EM algorithm is that it is easier to maximize the complete data likelihood than the marginal likelihood. Given initial values  $(\theta_0, \sigma_0^2)$ , the algorithm produces a sequence of estimates, that converge to their incomplete data maximum likelihood estimates (Dempster et al., 1977). Let  $(\tilde{\theta}_r, \tilde{\sigma}_r^2)$  be the current estimates of  $\theta$  and  $\sigma^2$  in the  $r^{\text{th}}$  EM iteration. The algorithm runs as follows for the  $(r+1)^{\text{th}}$  iteration. First, in the expectation (E-) step, the missing data,  $\boldsymbol{\Theta}^{(T)}$ , are

replaced with their expectation:

$$Q_{(r+1)}(\theta, \sigma^2 | \tilde{\theta}_r, \tilde{\sigma}_r^2) = \mathbb{E}_{\Theta^{(T)} | \mathbf{X}, \tilde{\theta}_r, \tilde{\sigma}_r^2} [f(\mathbf{X}, \Theta^{(T)} | \tilde{\Gamma}, \theta, \sigma^2)]. \quad (4.5)$$

The expectation in (4.5) is taken with respect to the unobserved  $\Theta^{(T)}$ , conditional on the observed data  $\mathbf{X}$  and the current estimates of the unknown parameters,  $\tilde{\theta}_r$  and  $\tilde{\sigma}_r^2$ . Based upon the definition of the complete data log-pseudolikelihood in (4.4),  $Q_{(r+1)}(\theta, \sigma^2 | \tilde{\theta}_r, \tilde{\sigma}_r^2)$  can be decomposed into two summations,

$$\begin{aligned} Q_{(r+1)}(\theta, \sigma^2 | \tilde{\theta}_r, \tilde{\sigma}_r^2) &= \sum_{t=1}^T \mathbb{E}_{\theta^{(t)} | \mathbf{X}, \tilde{\theta}_r, \tilde{\sigma}_r^2} \left\{ \log f(\theta^{(t)} | \theta, \sigma^2) \right\} \\ &+ \sum_{t=1}^T \sum_{l=1}^n \mathbb{E}_{\theta^{(t)} | \mathbf{X}, \tilde{\theta}_r, \tilde{\sigma}_r^2} \left\{ \log f(\mathbf{x}_l^{(t)} | \tilde{\gamma}^{(t)}, \theta^{(t)}) \right\}. \end{aligned} \quad (4.6)$$

To ease notation, henceforth, we suppress the subscript and conditional terms of all expectations in the E- and M-steps. As such, expectations will be taken over the distribution  $f(\theta^{(t)} | \mathbf{X}, \tilde{\theta}_r, \tilde{\sigma}_r^2)$ , unless otherwise stated.

The next step is the M-step which maximizes  $Q_{(r+1)}$  to yield new estimates:

$$(\tilde{\theta}_{(r+1)}, \tilde{\sigma}_{(r+1)}^2) = \arg \max_{\theta, \sigma^2} Q_{(r+1)}(\theta, \sigma^2 | \tilde{\theta}_r, \tilde{\sigma}_r^2).$$

Partial differentiation of (4.6) with respect to  $\theta_r$  and  $\sigma_r^2$  gives the following updated estimates,

$$\tilde{\theta}_{r+1} = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\theta^{(t)}] \quad (4.7a)$$

$$\tilde{\sigma}_{r+1}^2 = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[(\theta^{(t)} - \tilde{\theta}_{r+1})^2]. \quad (4.7b)$$

The above can be found by noting that the second summation in  $Q_{(r+1)}$  does not depend on either  $\theta$  or  $\sigma^2$  and that each  $\theta^{(t)}$  has density given by a Gaussian distribution with mean  $\tilde{\theta}_r$  and variance

$\tilde{\sigma}_r^2$  at the  $r^{th}$  iteration.

The Monte Carlo portion of the MCEM algorithm is introduced due to the fact that the expectation over  $\theta^{(t)}$  in the E-step cannot be directly computed, as the integration required to do so is intractable. We deal with the fact that we cannot directly compute these expectations by using a Markov Chain Monte Carlo (MCMC) sampling method, such as the Metropolis-Hastings algorithm. In the  $r^{th}$  iteration of the EM algorithm, we obtain a sample  $(\theta_{r1}^{(t)}, \theta_{r2}^{(t)}, \dots, \theta_{rS}^{(t)})$  from  $f(\theta^{(t)} | \mathbf{X}, \tilde{\theta}_r, \tilde{\sigma}_r^2)$  for  $t = 1 \dots, T$ , where  $S$  denotes the dependence on the MCMC sample size. Correspondingly, the Monte Carlo estimates of the expectations in (4.7) are given by

$$\hat{\mathbb{E}}[\theta^{(t)}] = \frac{1}{S} \sum_{s=1}^S \theta_{rs}^{(t)} \quad (4.8a)$$

$$\hat{\mathbb{E}}[(\theta^{(t)} - \tilde{\theta}_{r+1})^2] = \frac{1}{S} \sum_{s=1}^S (\theta_{rs}^{(t)} - \tilde{\theta}_{r+1})^2. \quad (4.8b)$$

If the Markov chain converges to its target distribution then,  $\hat{\mathbb{E}}[\{\theta^{(t)}\}^2] \xrightarrow{P} \mathbb{E}[\{\theta^{(t)}\}^2]$  and  $\hat{\mathbb{E}}[\theta^{(t)}] \xrightarrow{P} \mathbb{E}[\theta^{(t)}]$ .

Termination of the MCEM algorithm is not defined by usual convergence criteria used for the standard EM algorithm. This is due to stochastic nature of Monte Carlo portion of the algorithm. Instead, the average of the final  $g$  MCEM estimates are taken,

$$\tilde{\theta} = \frac{1}{g} \sum_{G=0}^{g-1} \tilde{\theta}_{R-G} \quad (4.9a)$$

$$\tilde{\sigma}^2 = \frac{1}{g} \sum_{G=0}^{g-1} \tilde{\sigma}_{R-G}^2, \quad (4.9b)$$

where  $R$  is the final iteration of the MCEM algorithm.

#### 4.2.5. A Monte Carlo likelihood ratio test

Suppose we are interested in determining if the two microbes have a pre-specified dependence structure given by  $\theta_0$ . We can test this by comparing the likelihoods under our two-stage maximum

likelihood estimate and the null parameter value. The likelihood ratio is given by,

$$\frac{L(\tilde{\theta}, \tilde{\sigma}^2)}{L(\theta_0, \tilde{\sigma}_0^2)} = \prod_{t=1}^T \frac{L_t(\tilde{\theta}, \tilde{\sigma}^2)}{L_t(\theta_0, \tilde{\sigma}_0^2)} = \prod_{t=1}^T \frac{f(\mathbf{x}^{(t)} | \tilde{\gamma}^{(t)}, \tilde{\theta}, \tilde{\sigma}^2)}{f(\mathbf{x}^{(t)} | \tilde{\gamma}^{(t)}, \theta_0, \tilde{\sigma}_0^2)}. \quad (4.10)$$

Typically, the two times the log of (4.10) has an asymptotically scaled  $\chi^2$  distribution (Deek and Li, 2021). As mentioned in the previous sections, direct calculation of these likelihoods are not feasible.

Therefore, we use the following Monte Carlo approximation. Using the properties of conditional probabilities and Bayes rules, it has been shown (Thompson, 1994) that Eq. 4.10 is equivalent to

$$\frac{L(\tilde{\theta}, \tilde{\sigma}^2)}{L(\theta_0, \tilde{\sigma}_0^2)} = \prod_{t=1}^T \mathbb{E}_0 \left[ \frac{f(\mathbf{x}^{(t)}, \theta^{(t)} | \tilde{\gamma}^{(t)}, \tilde{\theta}, \tilde{\sigma}^2)}{f(\mathbf{x}^{(t)}, \theta^{(t)} | \tilde{\gamma}^{(t)}, \theta_0, \tilde{\sigma}_0^2)} \middle| \mathbf{x}^{(t)} \right]. \quad (4.11)$$

The expectation in Eq. 4.11 is taken with respect to the distribution of  $\theta^{(t)}$  given  $\mathbf{x}^{(t)}$  and the null parameters  $\theta_0$  and  $\tilde{\sigma}_0^2$ . Finally, the Monte Carlo estimate of each of (4.11) can be formed using

$$\frac{L(\tilde{\theta}, \tilde{\sigma}^2)}{L(\theta_0, \tilde{\sigma}_0^2)} \approx \prod_{t=1}^T \frac{1}{S_0} \sum_{s=1}^{S_0} \frac{f(\mathbf{x}^{(t)}, \theta_s^{(t)} | \tilde{\gamma}^{(t)}, \tilde{\theta}, \tilde{\sigma}^2)}{f(\mathbf{x}^{(t)}, \theta_s^{(t)} | \tilde{\gamma}^{(t)}, \theta_0, \tilde{\sigma}_0^2)}, \quad (4.12)$$

where  $\theta_s^{(t)}$  is a MCMC realization sampled from  $f(\theta^{(t)} | \mathbf{x}^{(t)}; \tilde{\gamma}^{(t)}, \theta_0, \tilde{\sigma}_0^2)$ . Such an estimation works the best when  $\tilde{\theta}$  is close to  $\theta_0$ . We define the test based upon Eq. 4.12 as the Monte Carlo likelihood ratio test (mcLRT). Moreover, using the definition of conditional probability, the above ratio can be reduced to,

$$\frac{L(\tilde{\theta}, \tilde{\sigma}^2)}{L(\theta_0, \tilde{\sigma}_0^2)} \approx \prod_{t=1}^T \frac{1}{S_0} \sum_{s=1}^{S_0} \frac{f(\theta_s^{(t)} | \tilde{\theta}, \tilde{\sigma}^2)}{f(\theta_s^{(t)} | \theta_0, \tilde{\sigma}_0^2)}. \quad (4.13)$$

It is often the case that we are interested in testing for the independence of the two microbes, meaning the pair does not have a temporally conserved covariation. In this setting, the asymptotic distribution of the test statistic reduces to a  $\chi^2$  (Deek and Li, 2021). For the Frank copula, which is the parametric copula function we focus on,  $\theta_0 = 0$ .

### 4.3. Simulation studies

Simulation studies were used to assess the estimation accuracy of the two-stage Monte Carlo EM procedure. The data was simulated according to the following generative model, using the Rosenblatt transformation of the Frank copula function. We focus on the Frank copula function as it can model both positive and negative dependence, and the magnitude of dependence is symmetric around zero. First, define  $U = F_i(x_i)$ ,  $V = F_j(x_j)$  and

$$W = C_{v|u}(u, v) := \frac{\partial C(u, v)}{\partial u} = Pr(V = v|U = u).$$

The generative algorithm proceeds as follows:

---

**Algorithm 4** Data generation from a random-effects Frank copula with mixture margins.

---

- 1: Set  $\theta$  and  $\sigma^2$ .
  - 2: For every  $t = 1, \dots, T$  and  $l = 1, \dots, n$ :
  - 3: Draw  $\theta^{(t)} \sim N(\theta, \sigma^2)$ .
  - 4: Draw  $U_l^{(t)} \sim \text{Uniform}(0,1)$  and  $W_l^{(t)} \sim \text{Uniform}(0,1)$ .
  - 5: Solve for  $v_l^{(t)}$  using:  $v = -\frac{1}{\theta^{(t)}} \log\left\{1 + \frac{w(e^{-\theta^{(t)}} - 1)}{w + e^{-\theta^{(t)}}u(1-w)}\right\}$
  - 6: Solve for  $x_{il}^{(t)}$  using the definition of  $U$ :  $x_{il}^{(t)} = F_i^{-1}(u_l^{(t)}) = \begin{cases} 0 & \text{if } u_l^{(t)} \leq p_{li}^{(t)} \\ F_{beta}^{-1}\left(\frac{u_l^{(t)} - p_{li}^{(t)}}{1 - p_{li}^{(t)}}\right) & \text{if } u_l^{(t)} > p_{li}^{(t)} \end{cases}$
  - 7: Repeat Step (6) for  $x_{jl}^{(t)}$  and  $v_l^{(t)}$  is the same.
- 

The above process was repeated if (i) there were fewer non-zero relative abundances than parameters in the marginal regression models for either microbial taxon, (ii) the two taxa are mutually exclusive, i.e. if one taxon has a non-zero relative abundance the other must be absent, or if only one pair of observations has non-zero relative abundance for both taxa. This avoids the situation where the dependence parameter hits the lower boundary of estimation and/or causes unstable estimates.

Simulations were performed under varying parameters values for  $\theta$  and  $\sigma^2$  to evaluate the robustness of the estimation procedure. The parameters specified in simulation were selected based on what we would expect to see in real microbiome studies. Three simulation scenarios assumed intercept-only, or no covariate effect, models for the margins. For these models the zero-inflation probabilities were

set to  $(p_i, p_j) = (0.4, 0.5)$ . Additionally, the mean and dispersion parameters of the beta portion were set to  $(\mu_i, \mu_j) = (2/7, 3/9)$  and  $(\phi_i, \phi_j) = (7, 9)$ , respectively. Finally, the true dependence parameter was selected as either  $\theta = 0.5$  or 4 and standard deviation  $\sigma = 0.5$  or 1.3.

We can incorporate dependence between successive relative abundance measurements by using an autoregressive marginal model. Under this framework, an AR( $q$ ) model includes the relative abundance from the  $q$  previous time points as covariates in the generalized linear models for  $p$ ,  $\mu$  and/or  $\phi$ . We simulated data from an AR(1) structure for the zero-inflation and mean parameters using logistic regression models,  $\text{logit}(p_k^{(t)} | x_k^{(t-1)}) = \rho_{k0}^{(t)} + \rho_{k1}^{(t)} x_k^{(t-1)}$  and  $\text{logit}(\mu_k^{(t)} | x_k^{(t-1)}) = \delta_{k0}^{(t)} + \delta_{k1}^{(t)} x_k^{(t-1)}$ ,  $k = i, j$ . Here the subscript zero represents the model intercept and one denotes the regression coefficient for the relative abundance at the prior time point. The dispersion parameter was simulated from an intercept only model. We set the true parameter values as follows:  $\{(\rho_{i0}^{(t)}, \rho_{i1}^{(t)}), (\rho_{j0}^{(t)}, \rho_{j1}^{(t)})\} = (\{1, -4\}^\top, \{1, -4\}^\top)$ ,  $\{(\delta_{i0}^{(t)}, \delta_{i1}^{(t)}), (\delta_{j0}^{(t)}, \delta_{j1}^{(t)})\} = (\{-1, 1.5\}^\top, \{-1, 1.5\}^\top)$ ,  $(\kappa_{i0}^{(t)}, \kappa_{j0}^{(t)}) = (1, 1)$ . We also specify  $\theta = -3$ , and  $\sigma = 1$ .

Simulations were repeated 50 times, the number of time points was set to  $T = 25$ , and we used a sample size of  $n = 50$  and 100 for each of the four model settings. Moreover, for each simulated data set, the EM algorithm was run between 12-20 iterations and the Markov chains were run for a length of 2500-4000. Figure 4.2 shows that, in general across the 50 simulation runs,  $\tilde{\theta}$  is an empirically unbiased estimator for the large sample size ( $n = 100$ ). Similarly, our estimator  $\tilde{\sigma}$  performs reasonably well. The results are similar for the smaller sample size of  $n = 50$  (Figure C.1). Furthermore, the trace plots of the EM estimates at each iteration (not shown) indicate random variability of the estimates around the true value. These plots also show that the MCEM algorithm quickly stabilizes for the  $\tilde{\theta}$  estimates, while estimates of  $\tilde{\sigma}$  are slower to stabilize.

We further used simulations to assess the Type I error rate of the Monte Carlo likelihood ratio test for independence. To do so, data was simulated under the independence model ( $\theta = 0$  for a Frank copula). The remaining model parameters were specified as above, as well as the sample size and number of time points. The estimation results from 500 replicates were similar to the those discussed above under the alternative hypothesis for both sample sizes (Figures 4.2 and C.1). We

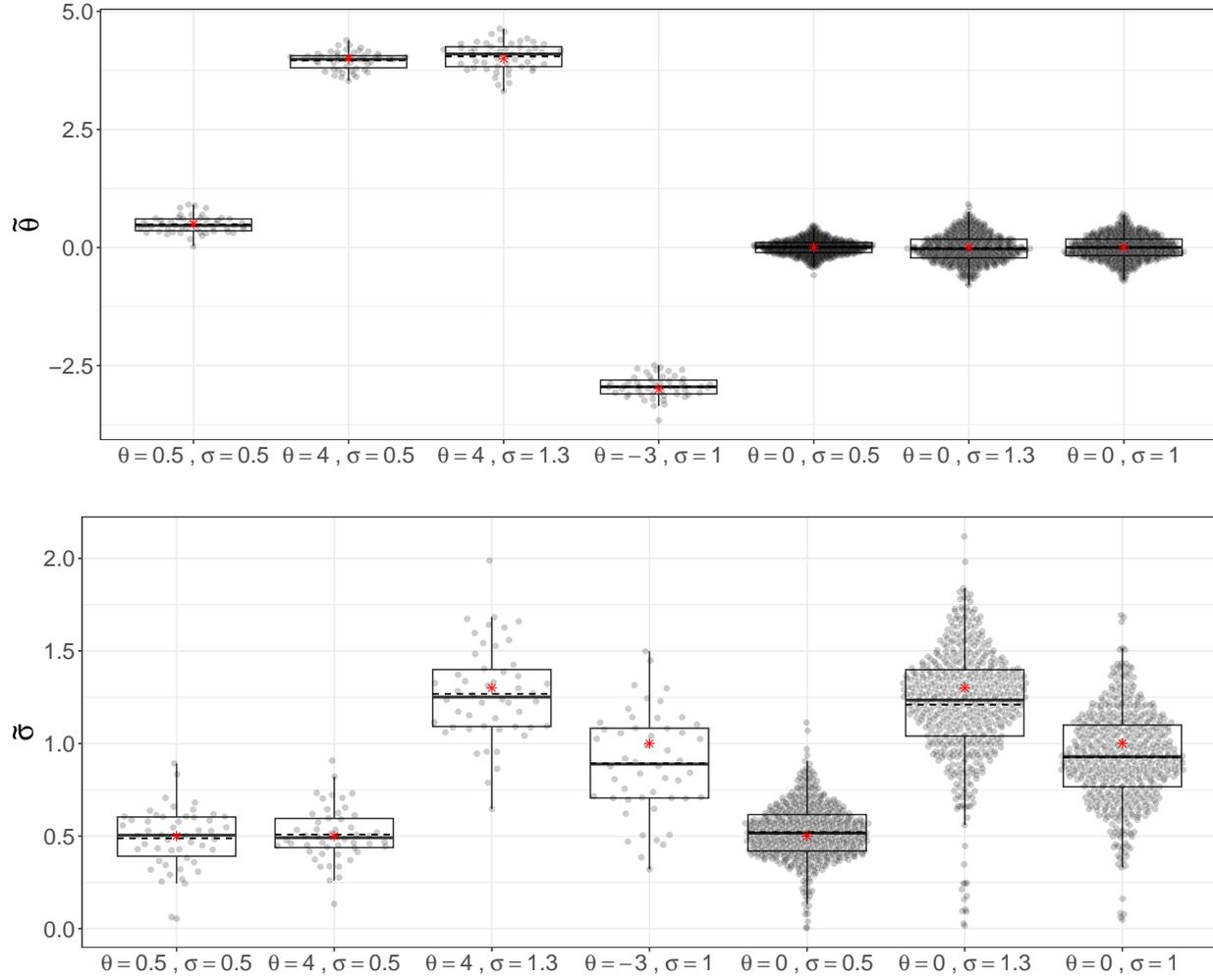


Figure 4.2: Boxplots of the  $\tilde{\theta}$  and  $\tilde{\sigma}$  estimates under seven different parameter ( $\theta$  and  $\sigma$ ) settings in simulation and  $n = 100$ . The first four are under the alternative hypothesis ( $\theta \neq 0$  for the Frank copula) and the last three are under the null of no conserved covariation structure ( $\theta = 0$ ). The black dashed line represents the mean across all runs. The red star indicates the true value specified under simulation.

find that the Type I error rate of the Monte Carlo LRT is generally well controlled at the nominal 0.05 level. Specifically, when  $n = 50$  the Type I error of the AR(0) models with  $\sigma = 0.05$  and 1.3 are 0.038 and 0.04, respectively. Likewise, the AR(1) model with  $\sigma = 1$  has a Type I error of 0.05. For the large sample size of  $n = 100$  the error is 0.046 ( $\sigma = 0.05$ ) and 0.062 ( $\sigma = 1.3$ ) for the AR(0) models, and 0.05 for the AR(1) model.

We further compared our proposed test to two naive methods. The first calculates a single Pearson correlation using all the data. The second, inspired by existing methods that use averaged time-specific metrics, calculates the Pearson correlation at each time point separately then averages them on the Fisher’s z-scale. By pooling the data, the first naive method resulted in inflated Type I error rates that range between 0.108 and 0.358. In contrast, the second method resulted in a conservative test, which rejects the null hypothesis in every simulation under the null. We did not compare to rank correlations (i.e. Spearman’s correlation, Kendall’s tau) as they have been shown to exhibit substantial bias, compared to copula based estimation procedures, in data with moderate to high zero-inflation (Deek and Li, 2021).

#### 4.4. Analysis of temporally conserved microbial networks during childhood development

##### 4.4.1. Conserved microbial covariation network estimation in children with and without antibiotic exposure

We applied the proposed method to the DIABIMMUNE antibiotics cohort from the Broad Institute (Yassour et al., 2016). The cohort consists of 39 children from Finland. The gut microbiome of each child is sampled monthly over the first 36 months after birth, thus making the data set particularly well suited for our method. These data were originally used to perform a natural history study exploring the dynamics of the gut microbiome of children before stabilization to a mature state (Yassour et al., 2016). The data is publicly available under NCBI BioProject ID PRJNA290381. In this analysis, we used the bacterial relative abundance data from the 16S sequencing.

Twenty children received between nine and fifteen antibiotic treatments, allowing for the quantification of the influence of antibiotic use on microbial covariation networks, in comparison to those

who had never received antibiotics. As a result we split the data set into two groups (antibiotics vs. no antibiotics) and performed estimation and testing of conserved covariation measures for each microbe pair. The study collected a total of 1002 samples, 483 of which are from children who received antibiotics and 519 from those who did not, with 105 unique genera. We removed any sequencing reads that were unassigned or ambiguously assigned at the genus level. Furthermore, we only used genera with at least a 10% prevalence across all samples in each group. This resulted in a total of 51 unique genera, from which 1275 pairs can be formed. For each pair, time points with less than 25% prevalence for either taxon were removed and any pair with less than ten time points remaining were removed. This left a total of 885 and 878 pairs for antibiotic and no antibiotic groups, respectively.

For all pairs, we performed MCEM estimation and testing for independence between the two microbes. We adjusted for multiple comparisons by controlling the false discovery rate of the Monte Carlo likelihood ratio test at 5% using the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001). A total of 112 and 201 pairs had a significant dependence parameter after FDR control in the antibiotics and no antibiotics groups, respectively. We compared the distribution of the estimated mean ( $\tilde{\theta}$ ) and standard deviation ( $\tilde{\sigma}$ ) of the random effect across significant and not significant pairs (Figure C.2). The distribution of  $\tilde{\theta}$  is shifted towards the null value of zero, as compared to that of non-significant pairs. Whereas the distribution of  $\tilde{\sigma}$  is similar regardless of significance, with most estimates falling around one and little evidence of small values close to zero. This implies that most of the taxa-pairs in real data fall into the no conserved covariation and conservation covariation scenarios ( $\sigma \neq 0$ ). Furthermore, there is little evidence of pairs with no covariation and strictly conserved covariation, which would be indicated by very small  $\tilde{\sigma}$  near zero.

These significant pairs were then used to construct conserved covariation networks of the microbial community in children with or without antibiotic exposure. To illustrate the conserved nature of these dependency structures, Figure 4.3 plots the posterior mean of the Monte Carlo sampled time-specific dependence parameters (i.e.  $\theta^{(t)}$ ) for a subset of pairs significant in both networks. Some pairs have little variability through time, as shown by their nearly flat horizontal lines. Other pairs

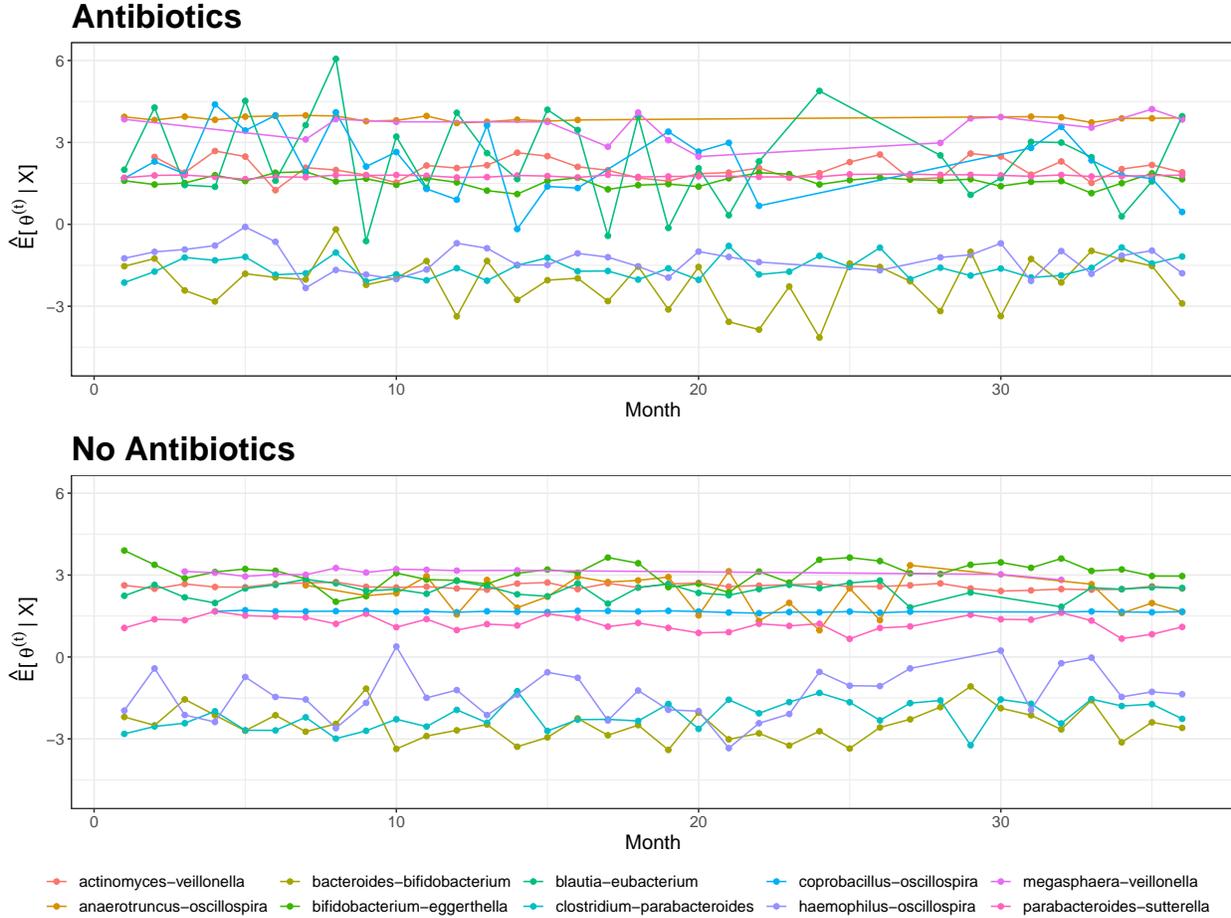


Figure 4.3: Plots of the posterior mean of  $\theta^{(t)}$  for a randomly selected subset of ten significant pairs, split by group (antibiotics vs. no antibiotics). Many of the pairs show relatively constant dependence parameters over time. Magnitude of the estimated posterior means and variability around the conserved value differ between groups for some pairs.

have larger variability through time, but are still centered around a conserved estimate.

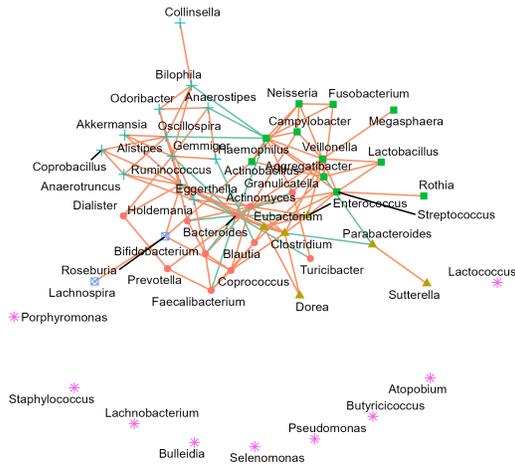
#### 4.4.2. Comparison to existing methods

Furthermore, we also applied the two naive methods described in the simulation studies section to the DIABIMMUNE data for comparison to the proposed method. The first method, that calculates a single Pearson correlation using all the day, detected 218 and 279 significant conserved covariations for antibiotics and no antibiotics groups, respectively. The second method, that calculates time specific Pearson correlations and averages them, detected 36 and 46 significant pairs for the two groups. These results are similar to those seen under simulation. The large number of pairs detected

by the first method is likely due to the inflated Type I error rate of its corresponding test, while the small number pairs detected by the second is due to the conservative nature of its test.

We also applied an existing method that aims to find an “ecogroup” of conserved covarying taxa to the data (Raman et al., 2019). The method recommends removing any taxon that was absent at any time point, to avoid excessive zeros in the time-specific covariance matrices and allow for direct computation of an average covariance matrix. This reduced the effective number of taxa by approximately half, to a total of 24 for the antibiotics group and 23 taxa for the no antibiotics group. Thresholding was done based upon the lower and upper 10% of the empirical distribution. We performed PCA on the resulting average covariance matrix, which is meant to capture of the conserved covariations, and find that PC1 captures 72% and 69% of the variability in the data for the antibiotics and no antibiotics groups, respectively. Using the top scores from PC1, we defined the ecogroup of bacteria that consistently covary with one another. From this ecogroup, a corresponding network was built such that two microbes have a conserved covariation (network edge) if at least one microbe belongs to the ecogroup and the pair’s average covariance is greater than the 80<sup>th</sup> percentile. This method identified 41 and 37 conserved taxa-pairs in the antibiotics and no antibiotics groups, respectively. Moreover, when we compared the networks from this method to those from our proposed method, we find that there are only 11 edges common in the antibiotics networks and 15 in the no antibiotics networks. We explored the effect of removing the ecogroup membership restriction on the Raman network, and find that the common edges nearly double for the antibiotics and no antibiotics analyses to 15 and 22. This is likely due to the fact that the majority of the conserved taxa-pairs our method detected are between two non-ecogroup microbes, specifically 87 pairs in the antibiotics and 147 pairs in the no antibiotics networks. This implies that the Raman method may be missing conserved covariations, that our method finds, by only building a network using ecogroup microbes. Additionally, we find the sparsity of the average covariance matrix and the resulting ecogroup is subject to change based upon the choice of thresholds.

### Antibiotics



### No Antibiotics

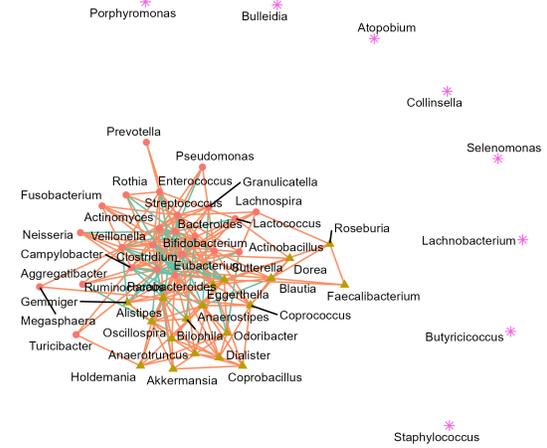


Figure 4.4: Conserved microbial covariation network diagrams for the antibiotics and no antibiotics groups. Networks were built using the estimated conserved mean dependence parameter ( $\tilde{\theta}$ ). Each node represents a bacterial genus, node shape and color correspond to graph cluster from an algorithm that maximized modularity. Two nodes share an edge if their FDR corrected p-value from the Monte Carlo likelihood ratio test was less than 0.05. Edge color corresponds to positive/negative (orange/green) dependence. The no antibiotics network has a higher edge density and lower modularity.

#### 4.4.3. Effects of antibiotic use on conserved covariation network properties

From the above copula model results, we built an adjacency matrix,  $\mathbf{A} = [a_{ij}]$ , and performed covariation network analysis. Two microbes are said to be connected ( $a_{ij} = 1$ ) if their mcLRT p-value was significant after FDR control, and are unconnected ( $a_{ij} = 0$ ) otherwise. Each of the 51 genera are represented by the nodes of the network diagram and edges correspond to the entries of the adjacency matrix (Figure 4.4). The graphs show that the overall network structures differ between the two groups, with only 69 common edges. The antibiotics network is less densely connected than the no antibiotics network, as seen by its higher mean distance (2.48 vs. 1.94), lower edge density (0.09 vs. 0.16), and lower cluster coefficient (0.32 vs. 0.40). The antibiotics network has five clusters, while the no antibiotics network has two clusters. Both networks consist of several unclustered taxa. Clusters were identified from a fast-greedy algorithm that aims to maximum network modularity (Clauset et al., 2004). Additionally, the antibiotics network is more modular (0.40 vs. 0.23) with many connections within clusters and few between. The adjusted Rand index quantifies the similarity between the partitions of set, and thus can be used to quantify the similarity between the clusters identified in the two networks. We observed the adjusted Rand index between the antibiotics and no antibiotics groups to be 0.296. We compared this to a null value, assuming no similarity in the clusters, from 1000 permutations that randomly permuted the cluster assignments for one network. The mean null Rand index is  $0.014 \pm 0.011$ , indicating the higher overlap between the two is more than expected by chance. Finally, we detected there to be a significant difference ( $p < 0.001$ ) in three- and four-node motif occurrence between the two networks, indicating there is also a difference in their local organizations.

We further compared the two networks using node specific centrality parameters. The antibiotics network has an average degree of 4.39 (sd=3.74), average closeness of 0.41 (sd=0.06), and average betweenness of 0.02 (sd=0.03). In contrast, the no antibiotics network has an average degree of 7.88 (sd=5.97), closeness of 0.53 (sd=0.07), and betweenness of 0.01 (sd=0.02). The lower average degree, closeness and betweenness of the antibiotics network compared to that of the no antibiotics provides further evidence of a less connected network. Node-wise comparisons of degree and closeness show

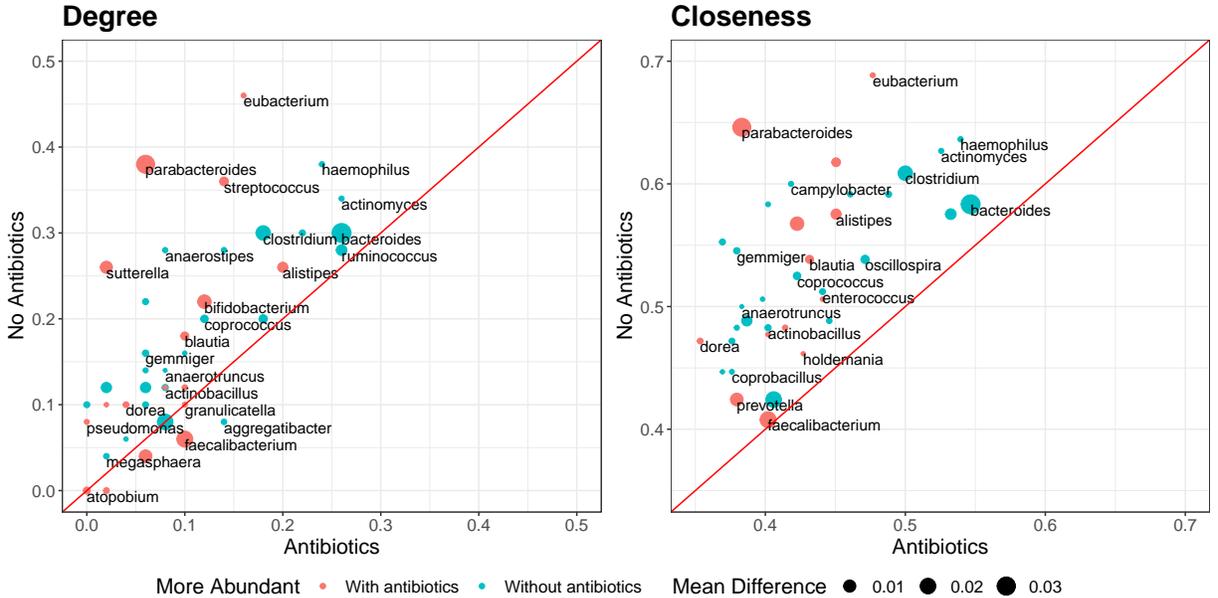


Figure 4.5: Node-level centrality statistics, degree and betweenness, plotted for each microbe to compare the antibiotics and no antibiotics networks. Dot size is proportional to the absolute difference in mean abundance between the two groups and node color indicates which group has larger mean abundance of the bacteria (i.e., blue indicates larger mean abundance in the no antibiotics group). Red line corresponds to  $y = x$ .

that there can be large differences in these network statistics for some microbes, and that these differences don't always align with differences in the microbe's mean abundance between the two groups (Figure 4.5). Thus indicating if analyses that focus on modeling the mean (e.g. regression) were used then differences in the gut microbiome's organization across antibiotics exposure status would likely be missed.

#### 4.4.4. Network robustness and stability

We also assessed the robustness and stability of the networks using targeted and random attacks (Figure 4.6). The 'attacks' involve sequentially removing the nodes from the network. Targeted attacks remove nodes in decreasing order of degree (i.e. nodes with higher degree are removed first) and random attacks remove nodes with equal probability. For each removed node, the diameter, defined as the length of the longest path, and global efficiency, the average of the shortest path between all pairs, are calculated.

Both networks had an immediate change in their diameter under targeted attack. Initially, both networks increased in their diameter, but the antibiotics network had a more rapid increase. Therefore implying a more fragile network under targeted attack for the antibiotics group. After 20-30% of the nodes had been removed, the two networks followed similar falling trajectories in their diameter, with that of the no antibiotics network shifted forward by five to eight nodes. Additionally, the efficiency of the two networks under targeted attack exhibited a similar decreasing pattern starting with the first node removal (results not shown), which is expected as we removed the most connected nodes first. Despite showing a similar pattern the efficiency of the no antibiotics network remained uniformly larger than that of the antibiotics network.

Under random attack, both networks showed resilience, with minimal changes in the efficiency until about 50% of the nodes were removed. At this point the efficiency of the antibiotics network began to decrease and that of the no antibiotics network began to increase. This indicates that both networks include redundant edges that help information be propagated through the network even as nodes are randomly removed. Furthermore, the efficiency of the no antibiotics network remained consistently larger than that of the antibiotics networks. All together these results indicate that the no antibiotics network is less dependent on individual nodes, thus making it more stable than the antibiotics network.

#### 4.5. Discussion

To fully understand the functional properties and organization of the microbiome, it is necessary to begin studying its emergent properties. Such properties cannot be identified from single species, “parts-of-a-whole” approaches, as they only appear from interactions between the parts of the system. We define the bivariate dependency structure between all pairs of taxa as the emergent properties of interest in this chapter. In particular, we focused on those dependence structures that are temporally conserved to understand the organizational stability of a microbiome.

To identify such covariation patterns based on longitudinal microbiome data, we have developed a mixture margin random-effects copula model by assuming the time-specific dependence parameter of bivariate copulas follow a Gaussian distribution. At each time point the observed data are

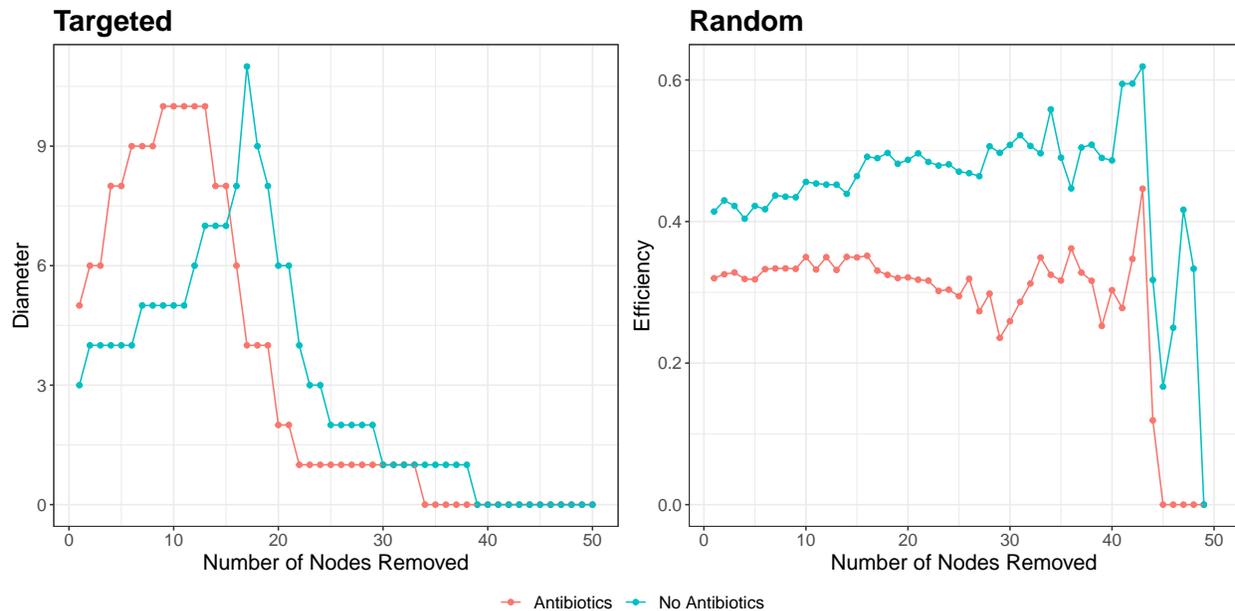


Figure 4.6: Network fragility measured by diameter and efficiency as nodes are sequentially removed in targeted and random attacks. The no antibiotics network is more robust to targeted attacks which remove nodes in decreasing degree order and maintains a higher efficiency than the antibiotics network in random attacks.

modeled with a bivariate copula, with a time-specific dependence parameter. Each of the time-sensitive parameters is assumed to be a random observation from a Gaussian distribution whose mean captures the conserved dependence structure. We developed an efficient Monte Carlo EM algorithm for estimation of the time-invariant dependence parameter and variance parameter, as well as a corresponding Monte Carlo likelihood ratio test.

The proposed mixture margin random-effects copula models are shown to be able to capture the conserved covariation relationship between two bacterial taxa based on longitudinal microbiome data. Such covariation relationships among the bacteria lead to construction of conserved covariation networks. Comparing such networks between two biological conditions can provide insights into how environmental factors, such as antibiotics use, affect the network structures. Our analysis of the DIABIMMUNE cohort data allow us to understand the impact of antibiotic use on the organizational structure of the infant microbiome. We find that antibiotic use leads to both global and local changes in the resulting microbial networks of conserved covariations. Furthermore, antibiotic use

is associated with a network that is less robust when subjected to network attacks.

We have focused on the estimation of pairwise covariations as they are useful when the sample size is small, as is the case in the DIABIMMUNE data. The proposed model can be used to assess the conditional pairwise associations by including all other taxa in the marginal models. Our model is developed to assess the conserved covariations between two taxa across time. As part of a future research project, it will be worthwhile to extend the method to allow for change points in covariations over time.

## CHAPTER 5

### DISCUSSION

This dissertation develops statistical methods, as well as computationally fast algorithms and open-source software, for modeling complex dependency structures in metagenomic sequencing data. These methods, algorithms, and software are built to capture the excessive zeros seen in such data. They also contribute to the growing emphasis on bivariate and multivariate models designed to elucidate how the microorganisms of a microbiome covary with one another. Chapter 1 introduced these novel methods. It detailed metagenomic sequencing studies and data, the importance of studying the microbiome’s influence on its host, and commonly used statistical techniques. It also discussed the current statistical and computational bottleneck in efficiently analyzing high-dimensional zero-inflated data from metagenomic sequencing technologies.

In Chapter 2, generative probabilistic mixture models for microbiome studies are introduced. When applied to metagenomic sequencing data, the latent variables of mixture models have important biological representation as the subcommunity structures that give rise to the observed counts data. Mixture models provide a way to simultaneously cluster microbes into subcommunities and characterize the prevalence of these subcommunities within the microbiome of subjects in a data set. Existing methodology, such as the Dirichlet-multinomial mixture model and Latent Dirichlet Allocation, as well as their limitations in microbiome data analysis are discussed. Accordingly, zero-inflated Latent Dirichlet Allocation (zinLDA), a multi-level hierarchical mixture model, is proposed. zinLDA uses a zero-inflated generalized Dirichlet distribution to describe the subcommunity specific taxa probabilities. This distribution introduces an additional set of latent variables that discriminates between structural and sampling zeros in each subcommunity. Distinguishing the type of zero count defines absent taxa versus taxa that suffer from dropout. Simulations show that zinLDA is more flexible and reduces over-smoothing without sacrificing computational efficiency. When applied to gut microbiome data from the American Gut Project zinLDA identified five microbial subcommunities with distinct bacterial compositions and finds relatively weak evidence of structural zeros in

the mostly healthy cohort.

Pairwise covariations between microbes are another type of dependency structure examined in this dissertation. Specifically, in Chapter 3, bivariate mixture margin copula models are proposed to estimate pairwise linear and non-linear covariation structures. The zero-beta mixture margins explicitly model the excessive zeros in the data with a point mass at zero, while the beta distribution is flexible in modeling proportion data. A primary advantage of a model-based approach is that it allows for adjustment of known confounders, such as host and environmental factors associated with the microbiome. Furthermore, a variance on the proposed estimator of the model's dependence parameter is derived providing an uncertainty measurement. Two stage estimation provides unbiased estimates of model parameters and the likelihood ratio test (LRT) for independence is uniformly more powerful than analogous tests based on sample correlation, as demonstrated in simulations. The estimated copula dependence parameters can be used to build biologically meaningful covariation networks. A covariation network built from a subset of healthy subjects from the American Gut Project has three clusters that tend to be dominated by different phylum and consist of mostly positive associations.

Chapter 4 detailed an extension of the above mixture margin copula model to the setting with longitudinal data. Longitudinal studies consist of repeated measures through time which allow for examination of the natural temporal variability in microbial covariations. This work focused on temporally conserved covariations as they are a marker of biological robustness and provide information about the stability of the ecosystem. Current methods estimate taxa covariance matrices to identify such covariations but suffer from instability and bias, as well as strict sparsity assumptions due to the data's high dimension. The model from Chapter 3 is advanced by treating the time-specific copula dependence parameters as random variables drawn from a Gaussian distribution, whose mean is the population level conserved covariation structure and variance controls the variability in dependence at each time around this conserved value. Simulation studies indicate that the Monte Carlo EM algorithm provides reasonable estimates of the model parameters and the proposed Monte Carlo LRT better controls the Type I error rate than naive methods using Pearson's correlation. The

method is applied to the antibiotics cohort of the publicly available DIABIMMUNE project. Two conserved covariation networks are built, stratified based on antibiotics exposure (ever vs. never), and show differences in both global and local patterns. The no antibiotics network is more densely connected and less reliant on any individual microbe when exposed to attacks than the antibiotics network, which is more modular with many connections within but few between clusters.

The primary application area of this dissertation is in metagenomic sequencing studies, but the methods are flexible enough to be applied to a wide range of data types. A natural expansion would be to examine their utility in single-cell genomics and spatial transcriptomics, as well as with multi-modal omics data. Particularly, in single-cell genomics the use of copula models to estimate an individual specific pairwise co-expression networks using the dependence parameters can be explored. By the same token, there are many new, open statistical questions of how to integrate multi-modal omics data sets into a single analysis. In this setting, copulas may be able to elucidate microbe-metabolite and microbe-host gene associations, thus helping address biological and ecological questions regarding the microbiome's functional capacity.

## APPENDIX A

### SUPPLEMENTARY MATERIALS FOR CHAPTER 2

#### A.1. Supplementary Data

##### A.1.1. Notation and terminology

The following defines the parameters and notation of the proposed zero-inflated Latent Dirichlet Allocation model:

- $\mathbf{z}^{(d)} = (z_{d1}, \dots, z_{dN})$ : a vector of subcommunity assignments for the  $d^{th}$  biological sample.  $z_{dn} = j$  indicates the  $n^{th}$  sequencing read in the  $d^{th}$  sample belongs to the  $j^{th}$  subcommunity. There are  $K$  underlying/latent subcommunity variables.
- $\boldsymbol{\theta}^{(d)} = (\theta_{d1}, \dots, \theta_{dK})$ : a vector of mixture probabilities for the  $d^{th}$  biological sample.  $\theta_{dj} = P(z = j | \boldsymbol{\theta}^{(d)})$  is the mixture probability for the  $j^{th}$  subcommunity in the sample.
- $\boldsymbol{\beta}^{(j)} = (\beta_{1j}, \dots, \beta_{(V-1)j})$  and  $\beta_{Vj} = 1 - \sum_{i=1}^{V-1} \beta_{ij}$ : a vector of taxon probabilities for the  $j^{th}$  subcommunity.  $\beta_{ij} = P(w^i | z = j, \boldsymbol{\beta})$  is the probability of observing the  $i^{th}$  taxon under the  $j^{th}$  subcommunity.
- $\mathbf{Q}^{(j)} = (Q_{1j}, \dots, Q_{(V-1)j})$ : a set of independent zero-inflated Beta random variables used to construct  $\boldsymbol{\beta}^{(j)}$ .
- $\boldsymbol{\alpha}^{(d)} = (\alpha_{d1}, \dots, \alpha_{dK})$ : the hyperparameter of the Dirichlet prior of  $\boldsymbol{\theta}^{(d)}$ .
- $\boldsymbol{\pi}^{(j)} = (\pi_{1j}, \dots, \pi_{(V-1)j})$ : a hyperparameter of the ZIGD specifying the probability of being a structural zero for the  $i^{th}$  taxon under the  $j^{th}$  subcommunity.
- $\boldsymbol{\Delta}^{(j)} = (\Delta_{1j}, \dots, \Delta_{(V-1)j})$ : a vector of indicator variables, where  $\Delta_{ij} = I(Q_{ij} = 0) = I(\beta_{ij} = 0)$  is an indicator function for structural zeros.
- $\mathbf{a}^{(j)} = (a_{1j}, \dots, a_{(V-1)j})$ : a hyperparameter on the ZIGD of  $\boldsymbol{\beta}^{(j)}$ .

- $\mathbf{b}^{(j)} = (b_{1j}, \dots, b_{(V-1)j})$  : a hyperparameter on the ZIGD of  $\beta^{(j)}$ .

All subsequent derivations assume that the hyperparameters  $\alpha^{(d)}, \pi^{(j)}, \mathbf{a}^{(j)}, \mathbf{b}^{(j)}$  are symmetric such that  $\alpha_{dj} = \alpha, \pi_{ij} = \pi, a_{ij} = a$ , and  $b_{ij} = b, \forall i, j$ .

### A.1.2. Model derivation and inference

#### Probability model

The zero-inflated Latent Dirichlet Allocation model is a probabilistic hierarchical model with the following specifications:

$$\begin{aligned}
 w_{dn} | z_{dn}, \beta^{(z_{dn})} &\sim \text{Multinomial}(\beta^{(z_{dn})}) \\
 \beta^{(z_{dn})} | \Delta &\sim \begin{cases} 0 & \text{if } \Delta_{i, z_{dn}} = 1 \\ \text{GD}(a, b) & \text{if } \Delta_{i, z_{dn}} = 0 \end{cases} \\
 \Delta | \pi &\sim \text{Ber}(\pi) \\
 z_{dn} | \theta^{(d)} &\sim \text{Multinomial}(\theta^{(d)}) \\
 \theta^{(d)} &\sim \text{Dirichlet}(\alpha)
 \end{aligned}$$

#### Posterior distribution

The hierarchical nature of the zinLDA model readily lends itself to a Bayesian inference framework, where the key inferential distribution is the posterior distribution. For zinLDA, the posterior is the joint distribution of the four latent variables, given the observed sequencing reads:  $P(\theta, \mathbf{z}, \beta, \Delta | \mathbf{w}; \alpha, \pi, a, b)$ . To find an analytical form of the posterior we must first have the joint distribution,  $P(\theta, \mathbf{z}, \beta, \Delta, \mathbf{w} | \alpha, \pi, a, b)$ , then marginalize to yield the normalizing constant,

$P(\mathbf{w}|\alpha, \pi, a, b)$ . The joint distribution can be found as follows:

$$\begin{aligned}
P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{w}|\alpha, \pi, a, b) &= P(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta})P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta}|\alpha, \pi, a, b) \\
&= \prod_{d=1}^D P(\boldsymbol{\theta}^{(d)}|\alpha)P(\boldsymbol{\beta}|\boldsymbol{\Delta}, a, b)P(\boldsymbol{\Delta}|\pi) \prod_{n=1}^N P(w_{dn}|z_{dn}, \boldsymbol{\beta})P(z_{dn}|\boldsymbol{\theta}^{(d)}) \\
&= \prod_{d=1}^D \text{Dir}(\alpha)P(\boldsymbol{\beta}|\boldsymbol{\Delta}, a, b)\text{Ber}(\pi) \prod_{n=1}^N \prod_{i=1}^V (\beta_{ij}\theta_j)^{w_{dn}^i}
\end{aligned}$$

$$\text{Where: } \boldsymbol{\beta}^{(j)}|\boldsymbol{\Delta}^{(j)}, a, b \sim \begin{cases} 0 & \text{if } \Delta_{ij} = 1 \\ \text{GD}(a, b) & \text{if } \Delta_{ij} = 0 \end{cases}$$

$$P(\boldsymbol{\beta}^{(j)}|\boldsymbol{\Delta}^{(j)}, a, b) = \mathbb{I}(\boldsymbol{\beta}_{\bar{U}_j} = 0)\text{GD}(\mathbf{a}_{U_j}, \mathbf{b}_{U_j})$$

$$P(\boldsymbol{\beta}|\boldsymbol{\Delta}, a, b)P(\boldsymbol{\Delta}|\pi) = P(\boldsymbol{\beta}, \boldsymbol{\Delta}|\pi, a, b) = P(\boldsymbol{\beta}|\pi, a, b) \sim \text{ZIGD}(\pi, a, b)$$

It should be noted that the probability equality statement:  $P(\boldsymbol{\beta}, \boldsymbol{\Delta}|\pi, a, b) = P(\boldsymbol{\beta}|\pi, a, b)$  holds since the addition of  $\boldsymbol{\Delta}$  does not give any additional information about  $\boldsymbol{\beta}$  (i.e.,  $\beta_{ij} = 0$  if and only if  $\Delta_{ij} = 1$ ). Moreover,  $w_{dn}^i$  is an indicator that the  $n^{\text{th}}$  sequencing read in the  $d^{\text{th}}$  sample corresponds to the  $i^{\text{th}}$  unique taxon. Furthermore, the GD density is given by:

$$GD(a, b) = \prod_{i=1}^{V-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \beta_{ij}^{(a-1)} (1 - \beta_{1j} - \dots - \beta_{ij})^{c_{ij}}$$

Where  $c_{ij} = b_{ij} - a_{(i+1)j} - b_{(i+1)j} = -a$  for  $i = 1, \dots, V-2$  and  $c_{(V-1)j} = b_{(V-1)j} - 1 = b-1$  under the assumption of symmetric of  $a$  and  $b$ .

$$P(\mathbf{z}, \boldsymbol{\Delta}|\mathbf{w}) = \frac{P(\mathbf{w}|\mathbf{z}, \boldsymbol{\Delta})P(\mathbf{z})P(\boldsymbol{\Delta}|\pi)}{\sum_{\boldsymbol{\Delta}} \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Delta})}$$

Since  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  only appear in  $P(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta})$  and  $P(\mathbf{z}|\boldsymbol{\theta})$ , respectively, the integration required to marginalize over the two can be done separately. Due to the conjugate prior property of both the zero-inflated generalized Dirichlet and the Dirichlet distributions, the marginalization results in

known compound probability distributions. We begin by marginalizing over  $\boldsymbol{\beta}$ :

$$\begin{aligned}
P(\mathbf{w}|\mathbf{z}, \boldsymbol{\Delta}) &= \int P(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta})P(\boldsymbol{\beta}|\boldsymbol{\Delta})d\boldsymbol{\beta} \\
&= \prod_{j=1}^K \int \text{Multinomial}(\boldsymbol{\beta}^{(j)})I(\boldsymbol{\beta}_{\bar{U}_j} = \mathbf{0})\text{GD}(\mathbf{a}_{U_j}, \mathbf{b}_{U_j})d\boldsymbol{\beta}^{(j)} \\
&= \prod_{j=1}^K \int \prod_{i=1}^V \beta_{ij}^{n_j^{(i)}} \prod_{l=1_j}^{L_j-1} \frac{1}{B(a, b)} \beta_{u_{lj}}^{a-1} (1 - \beta_{u_{1j}} - \dots - \beta_{u_{lj}})^{c_{u_{lj}}} d\boldsymbol{\beta}^{(j)} \\
&= \prod_{j=1}^K \int \prod_{l=1_j}^{L_j-1} \beta_{u_{lj}}^{n_j^{(u_l)}} \frac{1}{B(a, b)} \beta_{u_{lj}}^{a-1} (1 - \beta_{u_{1j}} - \dots - \beta_{u_{lj}})^{c_{u_{lj}}} d\boldsymbol{\beta}^{(j)} \\
&= \prod_{j=1}^K \prod_{i \in U_j} \frac{B(a_{ij}^{(z)}, b_{ij}^{(z)})}{B(a, b)} = \prod_{j=1}^K \prod_{i \in U_j} \frac{\Gamma(a_{ij}^{(z)})\Gamma(b_{ij}^{(z)})}{\Gamma(a_{ij}^{(z)} + b_{ij}^{(z)})} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \\
&= \prod_{j=1}^K \frac{B(\mathbf{a}_{U_j}^{(z)}, \mathbf{b}_{U_j}^{(z)})}{B(\mathbf{a}_{U_j}, \mathbf{b}_{U_j})}
\end{aligned}$$

Where we define  $n_j^{(i)}$  as the number of times the  $i^{\text{th}}$  taxa is assigned to the  $j^{\text{th}}$  subcommunity,  $a_{u_{lj}}^{(z)} = a + n_j^{(u_l)}$ , and  $b_{u_{lj}}^{(z)} = b + n_j^{(u_{l+1})} + \dots + n_j^{(u_{L_j})}$ . The ratio of the two beta functions is defined to be one for the last element of each  $U_j$ .

Likewise, we marginalize over  $\boldsymbol{\theta}$ :

$$\begin{aligned}
P(\mathbf{z}) &= \int P(\mathbf{z}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \prod_{d=1}^D \int \text{Multinomial}(\boldsymbol{\theta}^{(d)})\text{Dir}(\boldsymbol{\alpha})d\boldsymbol{\theta}^{(d)} \\
&= \prod_{d=1}^D \int \left( \prod_{j=1}^K \theta_{dj}^{m_j^{(d)}} \right) \frac{\Gamma(\sum_{j=1}^K \alpha)}{\prod_{j=1}^K \Gamma(\alpha)} \prod_{j=1}^K \theta_{dj}^{\alpha-1} d\boldsymbol{\theta}^{(d)} \\
&= \prod_{d=1}^D \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{j=1}^K \Gamma(m_j^{(d)} + \alpha)}{\Gamma(\sum_{j=1}^K m_j^{(d)} + \alpha)} \\
&= \left( \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \right)^D \prod_{d=1}^D \frac{\prod_{j=1}^K \Gamma(m_j^{(d)} + \alpha)}{\Gamma(m^{(d)} + K\alpha)}
\end{aligned}$$

Where we define  $m_j^{(d)}$  as the number of times the  $j^{\text{th}}$  subcommunity occurs in the  $d^{\text{th}}$  biological

sample.

### A.1.3. Gibbs sampling of $P(\mathbf{z}, \Delta | \mathbf{w})$

Since the posterior distribution  $P(\mathbf{z}, \Delta | \mathbf{w})$  cannot be directly computed we use Gibbs sampling to sequentially sample each  $z_{dn}$  and  $\Delta_{ij}$  individually conditional on all other  $\mathbf{z}$  and  $\Delta$ . These samples are taken as draws from the target posterior distribution, which can be used to approximate inferential quantities of interest. To do so, we must find the full conditional distributions:

1.  $P(z_{dn} = j | \mathbf{z}_{-n}, \mathbf{w}, \Delta)$

2.  $P(\Delta_{ij} = 1 | \Delta_{-i}, \mathbf{w}, \mathbf{z})$

#### Gibbs sampling of $P(z_{dn} = j | \mathbf{z}_{-n}, \mathbf{w}, \Delta)$

$$\begin{aligned}
 P(z_{dn}^{(i)} = j | \mathbf{z}_{-n}, \mathbf{w}, \Delta) &= \frac{P(z_{dn}^{(i)} = j, \mathbf{z}_{-n}, \mathbf{w}, \Delta)}{P(\mathbf{z}_{-n}, \mathbf{w}, \Delta)} = \frac{P(\mathbf{z}, \mathbf{w}, \Delta)}{P(\mathbf{z}_{-n}, \mathbf{w}, \Delta)} \\
 &= \frac{P(\mathbf{w} | \mathbf{z}, \Delta) P(\mathbf{z}) P(\Delta | \pi)}{P(\mathbf{w} | \mathbf{z}_{-n}, \Delta) P(\mathbf{z}_{-n}) P(\Delta | \pi)} \\
 &= \frac{P(\mathbf{w} | \mathbf{z}, \Delta) P(\mathbf{z})}{P(\mathbf{w} | \mathbf{z}_{-n}, \Delta) P(\mathbf{z}_{-n})} \\
 P(z_{dn}^{(i)} = j | \mathbf{z}_{-n}, \mathbf{w}, \Delta) &= \begin{cases} \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{1j} \\ \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \prod_{t < i, t \in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } u_{1j} < i < u_{Lj} \\ \prod_{t < i, t \in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{Lj} \\ 0 & \text{if } i \notin U_j \end{cases}
 \end{aligned}$$

Gibbs sampling of  $P(\Delta_{ij} = 1 | \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z})$

$$\begin{aligned}
P(\Delta_{ij} = 1 | \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z}) &= \frac{P(\Delta_{ij} = 1, \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z})}{P(\mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z})} \\
&= \frac{P(\mathbf{w} | \mathbf{z}, \Delta_{ij} = 1, \mathbf{\Delta}_{-i}) P(\mathbf{z}) P(\Delta_{ij} = 1, \mathbf{\Delta}_{-i})}{P(\mathbf{w} | \mathbf{z}, \mathbf{\Delta}_{-i}) P(\mathbf{z}) P(\mathbf{\Delta}_{-i})} \\
&= \frac{P(\mathbf{w} | \mathbf{z}, \Delta_{ij} = 1, \mathbf{\Delta}_{-i}) P(\Delta_{ij} = 1, \mathbf{\Delta}_{-i})}{P(\mathbf{w} | \mathbf{z}, \mathbf{\Delta}_{-i}) P(\mathbf{\Delta}_{-i})} \\
&\propto P(\mathbf{w} | \mathbf{z}, \Delta_{ij} = 1, \mathbf{\Delta}_{-i}) P(\Delta_{ij} = 1, \mathbf{\Delta}_{-i}) \\
&= \begin{cases} 0 & \text{if } n_j^{(i)} > 0 \\ \pi_{ij} \prod_{k=1}^K \left\{ \prod_{l \in U_{j,-i}} \frac{B(a_{lk}^{(z)}, b_{lk}^{(z)})}{B(a,b)} \right\} \left\{ \sum_{\Delta} \prod_{l=1,-i}^{V-1} \pi_{lk}^{\Delta_{lk}} (1-\pi_{lk})^{(1-\Delta_{lk})} \right\} & \text{if } n_j^{(i)} = 0 \end{cases}
\end{aligned}$$

$$\begin{aligned}
P(\Delta_{ij} = 0 | \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z}) &\propto P(\mathbf{w} | \mathbf{z}, \Delta_{ij} = 0, \mathbf{\Delta}_{-i}) P(\Delta_{ij} = 0, \mathbf{\Delta}_{-i}) \\
&= (1-\pi_{ij}) \frac{B(a_{ij}^{(z)}, b_{ij}^{(z)})}{B(a,b)} \prod_{k=1}^K \left\{ \prod_{l \in U_{j,-i}} \frac{B(a_{lk}^{(z)}, b_{lk}^{(z)})}{B(a,b)} \right\} \left\{ \sum_{\Delta} \prod_{l=1,-i}^V \pi_{lk}^{\Delta_{lk}} (1-\pi_{lk})^{(1-\Delta_{lk})} \right\}
\end{aligned}$$

Putting these two results together gives:

$$P(\Delta_{ij} = 1 | \mathbf{\Delta}_{-i}, \mathbf{w}, \mathbf{z}) = \begin{cases} 0 & \text{if } n_j^{(i)} > 0 \\ \frac{\pi_{ij}}{\pi_{ij} + (1-\pi_{ij}) \frac{B(a_{ij}^{(z)}, b_{ij}^{(z)})}{B(a,b)}} & \text{if } n_j^{(i)} = 0 \end{cases}$$

#### A.1.4. Estimating $\beta$ and $\theta$

Both  $\beta$  and  $\theta$  can be estimated using the predictive distribution over taxon and subcommunity assignments of new sequencing reads. The estimator  $\hat{\beta}_{ij}$  is derived by:

$$\begin{aligned}
\hat{\beta}_{ij} &= Pr(w_{d,new}^{(i)} | z_{d,new}^{(i)} = j, \mathbf{w}, \mathbf{z}, \Delta) \\
&= \int \underbrace{Pr(w_{d,new}^{(i)} | \beta, z_{d,new}^{(i)} = j)}_{\beta_{ij}} \underbrace{Pr(\beta | \mathbf{w}, \mathbf{z}, \Delta)}_{\text{posterior}} d\beta \\
&= \int \beta_{ij} \prod_{l=1_j}^{L_j-1} \frac{1}{B(a_{ij}^{(z)}, b_{ij}^{(z)})} \beta_{u_{1j}}^{a_{u_{1j}}^{(z)}-1} (1 - \beta_{u_{1j}} - \dots - \beta_{u_{L_j}})^{c_{u_i}} d\beta^{(j)} \\
\hat{\beta}_{ij} &= \begin{cases} \frac{a+n_j^{(i)}}{a+n_j^{(i)}+b_{ij}^{(z)}} & \text{if } i = u_{1j} \\ \frac{a+n_j^{(i)}}{a+n_j^{(i)}+b_{ij}^{(z)}} \prod_{t < i, t \in U_j} \frac{b_{tj}^{(z)}}{a+n_j^{(t)}+b_{tj}^{(z)}} & \text{if } u_{1j} < i < u_{L_j} \\ \prod_{t < i, t \in U_j} \frac{b_{tj}^{(z)}}{a+n_j^{(t)}+b_{tj}^{(z)}} & \text{if } i = u_{L_j} \\ 0 & \text{if } i \notin U_j \end{cases}
\end{aligned}$$

Note that the estimate of  $\beta_{ij}$  does not include any function of  $\pi$ . Any function of  $\pi$  is only introduced via the density of  $\Delta$ ,  $Pr(\Delta | \pi)$ , which is not needed here since the predictive probability is conditional of  $\Delta$ . Next, the estimator  $\hat{\theta}_{ij}$  is derived by:

$$\begin{aligned}
\hat{\theta}_{dj} &= P(z_{d,new} = j | \mathbf{w}, \mathbf{z}) \\
&= \int \underbrace{Pr(z_{d,new} = j | \theta^{(d)})}_{\theta_{dj}} \underbrace{Pr(\theta^{(d)} | \mathbf{z})}_{\text{posterior}} d\theta^{(d)} \\
&= \int \theta_{dj} \frac{\Gamma(m_j^{(d)} + K\alpha)}{\prod_{j=1}^K \Gamma(m_j^{(d)} + \alpha)} \prod_{k=1}^K \theta_{dk}^{m_k^{(d)} + \alpha - 1} d\theta^{(d)} \\
&= \frac{\Gamma(m_j^{(d)} + K\alpha)}{\Gamma(m_j^{(d)} + \alpha)} \frac{\Gamma(m_j^{(d)} + \alpha + 1)}{\Gamma(m_j^{(d)} + K\alpha + 1)} \\
\hat{\theta}_{dj} &= \frac{m_j^{(d)} + \alpha}{m_j^{(d)} + K\alpha}
\end{aligned}$$

#### A.2. Supplementary figures

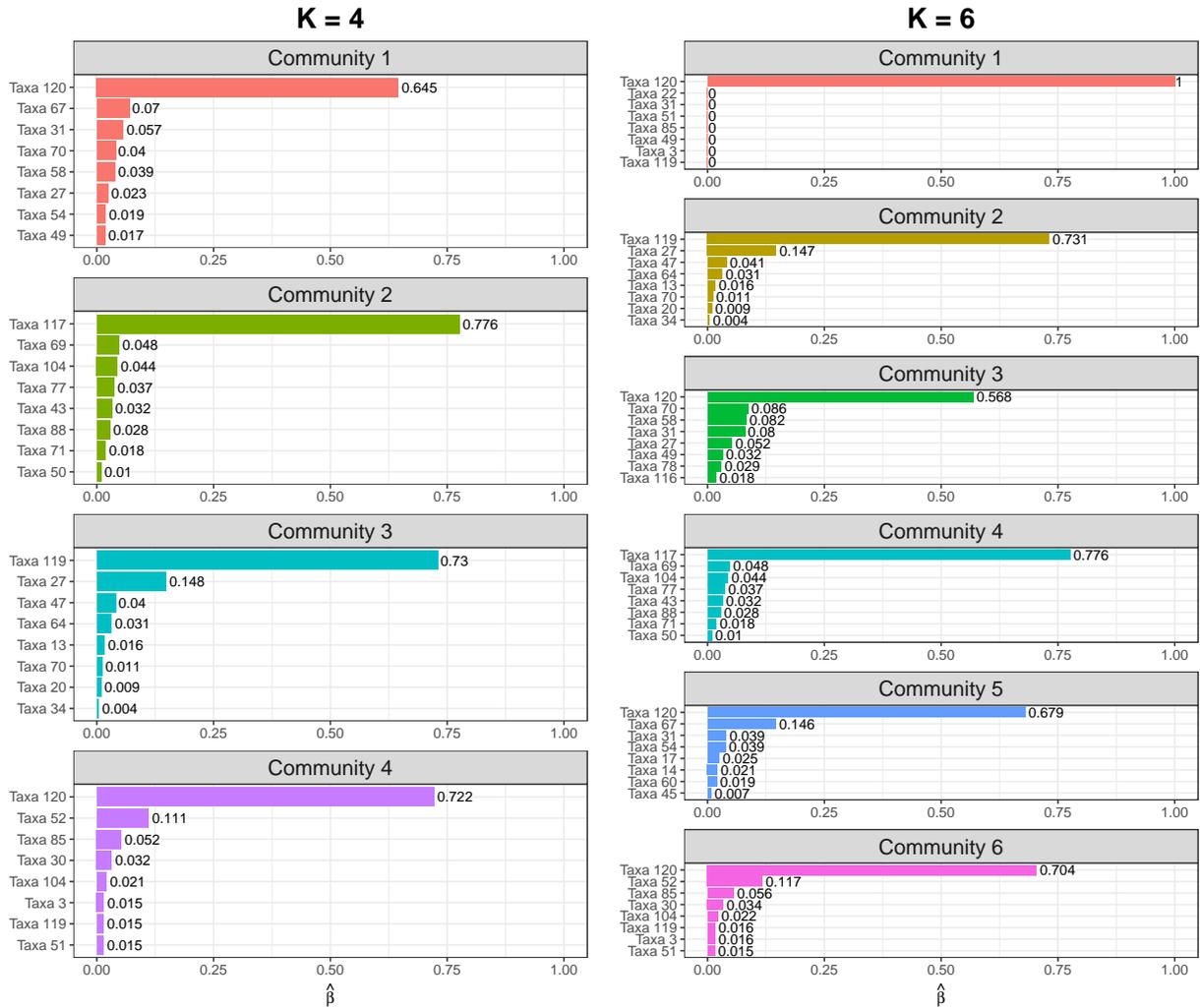


Figure A.1: Bar graphs of the top eight taxa for each subcommunity with their corresponding  $\hat{\beta}_{ij}$  values under model misspecification. Data was simulated under a true zero-inflated latent Dirichlet allocation model with five communities and observed  $V = 87$ . An underspecified model with four (left) and an overspecified model with six (right) subcommunities were fit to the data.

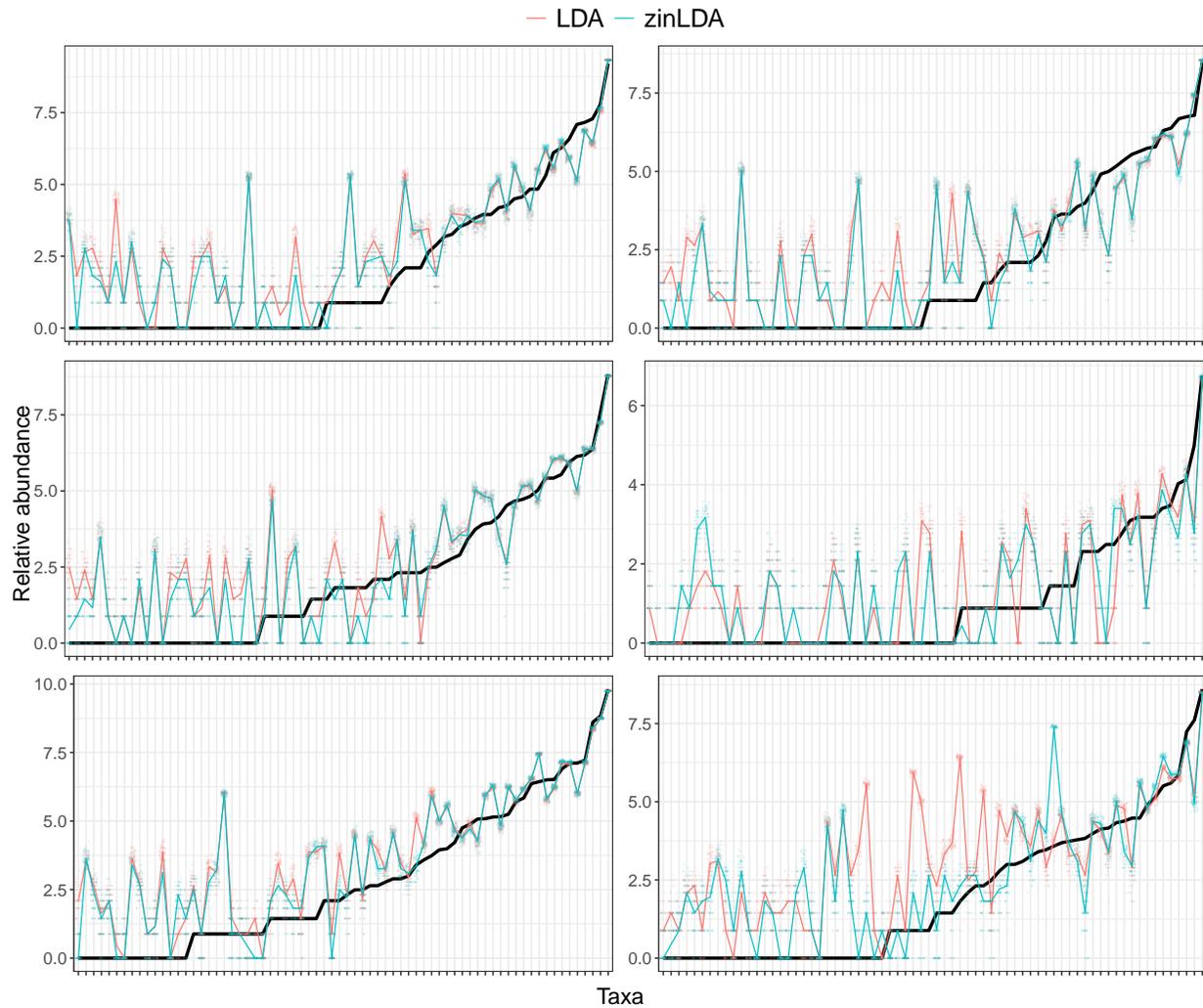


Figure A.2: Observed and posterior predictive simulated  $asinh$ -transformed taxon counts plotted in order of increasing observed abundance from the American Gut Project. Each panel is a different biological sample. The solid black line represents the observed counts. The pink and blue points are the counts from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median counts across the 50 data sets.

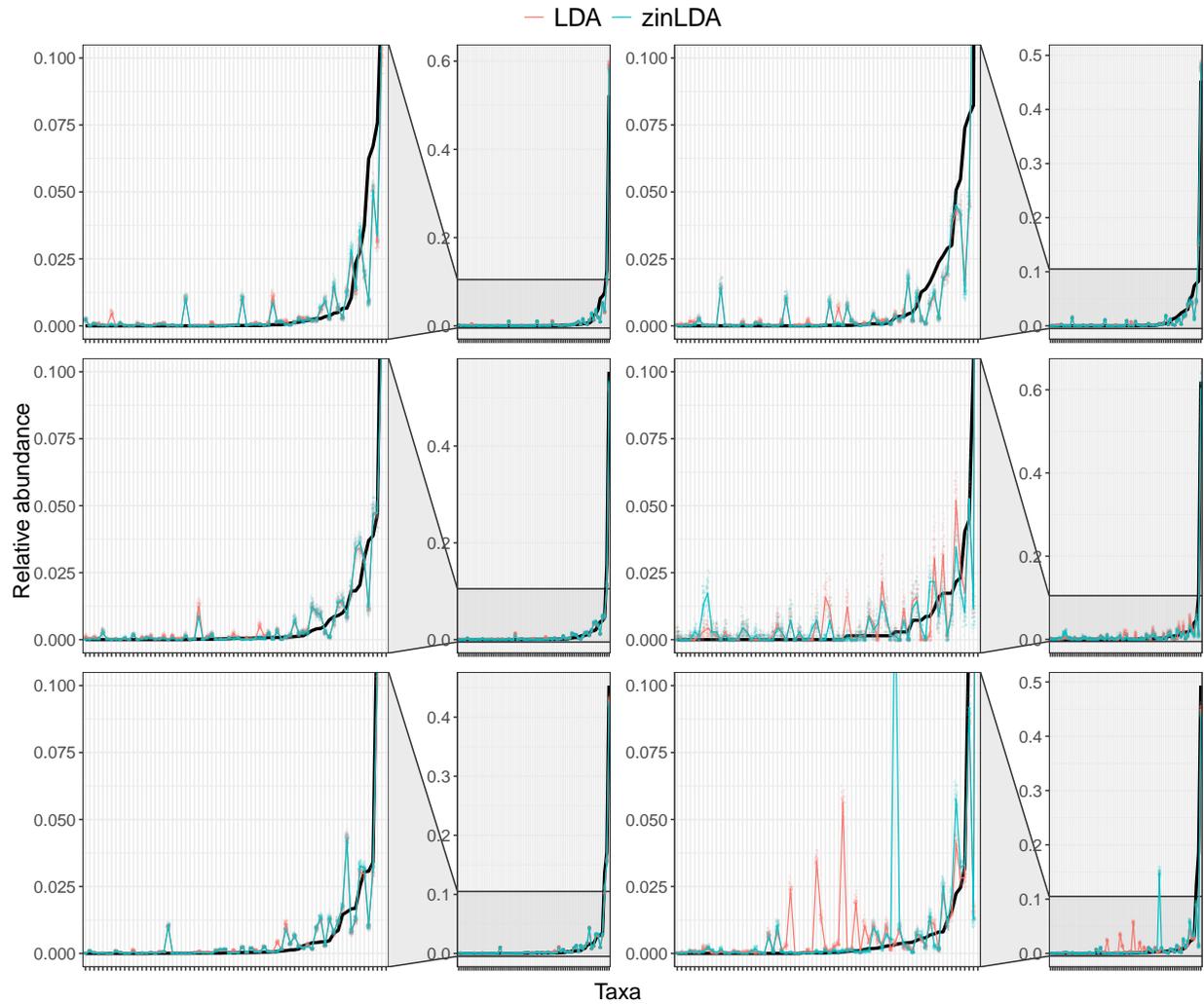


Figure A.3: Observed and posterior predictive simulated relative abundances of taxa from the same six samples selected in Figure A.2, plotted in order of increasing observed relative abundance from the American Gut Project. The solid black line represents the observed relative abundance. The pink and blue points are the relative relative abundances from 50 posterior predictive simulated data sets from the LDA and zinLDA models, respectively. The pink and blue solid lines represents the median abundances across the 50 data sets. The first and third column are zoom in on columns two and four, respectively, showing relative abundances between 0 and 0.1 only.

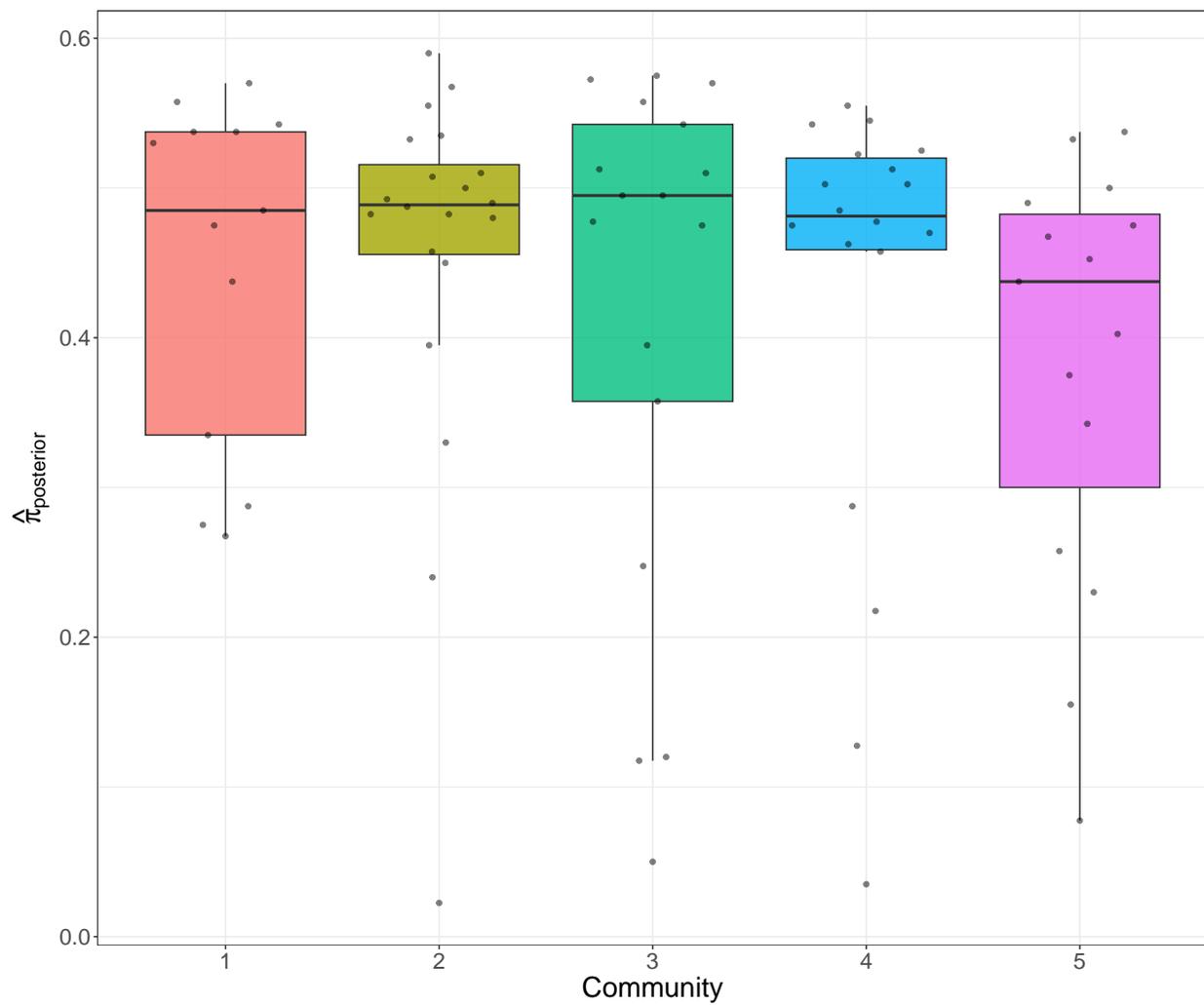


Figure A.4: Boxplot of the posterior estimates of  $\pi_{ij}$  for taxa with zero counts, split by subcommunity, from the zinLDA model applied to a subset of 1000 subjects from the American Gut Project.

## APPENDIX B

### SUPPLEMENTARY MATERIALS FOR CHAPTER 3

#### B.1. Supplementary data

##### B.1.1. Copula joint density function

For our specified zero-beta mixture margins, we can define four different joint densities based upon which component(s) of the pair are from the discrete zero portion and differentiation of the copula distribution function. Recall from the manuscript that  $\mathcal{M} = \{i, j\}$  is the index set,  $\mathcal{C}(\mathbf{x})$  contains the indices of  $\mathbf{x} = \{x_i, x_j\}$  with continuous  $F$  at  $x$ , and  $\mathcal{D}(\mathbf{x}) = \mathcal{M} - \mathcal{C}(\mathbf{x})$  is the set of indices of  $\mathbf{x}$  for which  $F$  has a jump point at  $x$ . We now define the four possible joint densities as follows:

- **S1:**  $x_i \neq 0, x_j \neq 0, \implies \mathcal{C} = \{i, j\}, \mathcal{D} = \emptyset$

$$f(x_i, x_j) = c(F_i(x_i), F_j(x_j))f_i(x_i)f_j(x_j)$$

- **S2:**  $x_i = 0, x_j \neq 0, \implies \mathcal{C} = \{j\}, \mathcal{D} = \{i\}$

$$\begin{aligned} f(x_i, x_j) &= f_j(x_j) \Delta_{F_i(x_i^-)}^{F_i(x_i)} C_{i|j}(\cdot | F_j(x_j)) \\ &= f_j(x_j) \{C_{i|j}(F_i(x_i) | F_j(x_j)) - C_{i|j}(F_i(x_i^-) | F_j(x_j))\} \\ &= f_j(x_j) \{C_{i|j}(p_i | F_j(x_j)) - C_{i|j}(0 | F_j(x_j))\} \\ &= f_j(x_j) C_{i|j}(p_i | F_j(x_j)) \end{aligned}$$

- **S3:**  $x_i \neq 0, x_j = 0, \implies \mathcal{C} = \{i\}, \mathcal{D} = \{j\}$

$$\begin{aligned}
f(x_i, x_j) &= f_i(x_i) \Delta_{F_j(x_j^-)}^{F_j(x_j)} C_{j|i}(\cdot | F_i(x_i)) \\
&= f_i(x_i) \{C_{j|i}(F_j(x_j) | F_i(x_i)) - C_{j|i}(F_j(x_j^-) | F_i(x_i))\} \\
&= f_i(x_i) \{C_{j|i}(p_j | F_i(x_i)) - C_{j|i}(0 | F_i(x_i))\} \\
&= f_i(x_i) C_{j|i}(p_j | F_i(x_i))
\end{aligned}$$

- **S4:**  $x_i = 0, x_j = 0, \implies \mathcal{C} = \emptyset, \mathcal{D} = \{i, j\}$

$$\begin{aligned}
f(x_i, x_j) &= \Delta_{F_i(x_i^-)}^{F_i(x_i)} \Delta_{F_j(x_j^-)}^{F_j(x_j)} C(\cdot) \\
&= \Delta_{F_i(x_i^-)}^{F_i(x_i)} C(\cdot, F_j(x_j)) - C(\cdot, F_j(x_j^-)) \\
&= C(F_i(x_i), F_j(x_j)) - C(F_i(x_i), F_j(x_j^-)) \\
&\quad - C(F_i(x_i^-), F_j(x_j)) + C(F_i(x_i^-), F_j(x_j^-)) \\
&= C(p_i, p_j) - C(p_i, 0) - C(0, p_j) + C(0, 0) \\
&= C(p_i, p_j)
\end{aligned}$$

### B.1.2. Score equation of the dependence parameter

Score equation with respect to dependence parameter  $\theta$  is given by:

$$\begin{aligned}
\tilde{U}_\theta &= \frac{n_1}{\theta} - \frac{n_1 e^{-\theta}}{e^{-\theta} - 1} - \sum_{i \in S_1} (\tilde{u} + \tilde{v}) - 2 \sum_{i \in S_1} \frac{-(\tilde{u} + \tilde{v})e^{-\theta(\tilde{u}+\tilde{v})} + \tilde{u}e^{-\theta\tilde{u}} + \tilde{v}e^{-\theta\tilde{v}} - e^{-\theta}}{e^{-\theta(\tilde{u}+\tilde{v})} - e^{-\theta\tilde{u}} - e^{-\theta\tilde{v}} + e^{-\theta}} \\
&+ \sum_{i \in S_2} \frac{-\tilde{p}_i e^{-\theta\tilde{p}_i}}{e^{-\theta\tilde{p}_i} - 1} - \sum_{i \in S_2} \tilde{v} - \sum_{i \in S_2} \frac{-(\tilde{p}_i + \tilde{v})e^{-\theta(\tilde{p}_i+\tilde{v})} + \tilde{p}_i e^{-\theta\tilde{p}_i} + \tilde{v}e^{-\theta\tilde{v}} - e^{-\theta}}{e^{-\theta(\tilde{p}_i+\tilde{v})} - e^{-\theta\tilde{p}_i} - e^{-\theta\tilde{v}} + e^{-\theta}} \\
&+ \sum_{i \in S_3} \frac{-\tilde{p}_j e^{-\theta\tilde{p}_j}}{e^{-\theta} - 1} - \sum_{i \in S_3} \tilde{u} - \sum_{i \in S_3} \frac{-(\tilde{u} + \tilde{p}_j)e^{-\theta(\tilde{u}+\tilde{p}_j)} + \tilde{u}e^{-\theta\tilde{u}} + \tilde{p}_j e^{-\theta\tilde{p}_j} - e^{-\theta}}{e^{-\theta(\tilde{u}+\tilde{p}_j)} - e^{-\theta\tilde{u}} - e^{-\theta\tilde{p}_j} + e^{-\theta}} \\
&- \frac{n_4}{\theta} + \sum_{i \in S_4} \frac{-(\tilde{p}_i + \tilde{p}_j)e^{-\theta(\tilde{p}_i+\tilde{p}_j)} + \tilde{p}_i e^{-\theta\tilde{p}_i} + \tilde{p}_j e^{-\theta\tilde{p}_j}}{\log \left\{ 1 + \frac{e^{-\theta(\tilde{p}_i+\tilde{p}_j)} - e^{-\theta\tilde{p}_i} - e^{-\theta\tilde{p}_j} + 1}{e^{-\theta} - 1} \right\} \left( 1 + \frac{e^{-\theta(\tilde{p}_i+\tilde{p}_j)} - e^{-\theta\tilde{p}_i} - e^{-\theta\tilde{p}_j} + 1}{e^{-\theta} - 1} \right)} \\
&+ \sum_{i \in S_4} \frac{e^{-\theta(\tilde{p}_i+\tilde{p}_j)+1} - e^{-\theta(\tilde{p}_i+1)} - e^{-\theta(\tilde{p}_j+1)} + e^{-\theta}}{\log \left\{ 1 + \frac{e^{-\theta(\tilde{p}_i+\tilde{p}_j)} - e^{-\theta\tilde{p}_i} - e^{-\theta\tilde{p}_j} + 1}{e^{-\theta} - 1} \right\} \left( 1 + \frac{e^{-\theta(\tilde{p}_i+\tilde{p}_j)} - e^{-\theta\tilde{p}_i} - e^{-\theta\tilde{p}_j} + 1}{e^{-\theta} - 1} \right)} (e^{-\theta} - 1)^2
\end{aligned}$$

### B.1.3. Proof of Theorem 1

*Proof.* For simplicity, we write  $\ell(\theta) = \ell(\theta, \tilde{\gamma}_i, \tilde{\gamma}_j)$ . By Taylor expansion, we have

$$\ell(\theta_0) = \ell(\tilde{\theta}) + (\theta_0 - \tilde{\theta})\ell'(\tilde{\theta}) + \frac{1}{2}(\theta_0 - \tilde{\theta})^2\ell''(\tilde{\theta}) + \dots$$

Since  $\tilde{\theta}$  is the value that maximizes  $\ell(\theta, \tilde{\gamma}_i, \tilde{\gamma}_j)$ , we have  $\ell'(\tilde{\theta}) = 0$  and

$$\Lambda = -2[\ell(\theta_0) - \ell(\tilde{\theta})] \asymp -(\theta_0 - \tilde{\theta})^2\ell''(\tilde{\theta}) = -\frac{n(\tilde{\theta} - \theta_0)^2}{v} \frac{\ell''(\tilde{\theta})v}{n}.$$

where  $v = \mathbf{V}_{7,7}$  is the  $(7, 7)^{th}$  entry of the covariance matrix of  $\tilde{\boldsymbol{\eta}}$ , which can be calculated as:

$$v = \mathcal{I}_{\theta\theta}^{-1} + \mathcal{I}_{\theta\theta}^{-2}(\mathcal{I}_{\theta 1}\mathcal{J}_{11}^{-1}\mathcal{I}_{1\theta} + \mathcal{I}_{\theta 2}\mathcal{J}_{22}^{-1}\mathcal{I}_{2\theta} + \mathcal{I}_{\theta 1}\mathcal{J}_{11}^{-1}\mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{I}_{2\theta} + \mathcal{I}_{\theta 2}\mathcal{J}_{22}^{-1}\mathcal{J}_{21}\mathcal{J}_{11}^{-1}\mathcal{I}_{1\theta}). \quad (\text{B.1})$$

Note that  $\frac{n(\tilde{\theta} - \theta_0)^2}{v} \xrightarrow{D} \chi_1^2$ . Thus, it suffices to deal with the ratio  $\frac{-\ell''(\tilde{\theta})v}{n}$ . Now since

$$-n^{-1}\ell''(\theta) = -n^{-1}\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{1}{n} \sum_{l=1}^n \frac{\partial^2 \log f(\mathbf{X}_l; \tilde{\gamma}_i, \tilde{\gamma}_j, \theta)}{\partial \theta^2},$$

by the Mean Value Theorem and the Law of Large Numbers,

$$-\ell''(\tilde{\theta})/n \xrightarrow{P} -\ell''(\theta)/n \xrightarrow{P} -\mathbb{E}[\ell''(\theta)] = \mathcal{I}_{\theta\theta},$$

which can be approximated by  $\tilde{\mathcal{I}}_{\theta\theta}$  using numerical methods and  $v$  can be estimated by a consistent estimator,  $\tilde{v}$ , such as a jackknife estimate.

We now define the following two-stage LRT statistic:

$$\Lambda' = (\tilde{v}\tilde{\mathcal{I}}_{\theta\theta})^{-1}\Lambda = \tilde{\omega}\Lambda \tag{B.2}$$

The above discussion implies

$$\Lambda' \rightarrow_D \chi_1^2, \quad \text{as } n \rightarrow \infty. \tag{B.3}$$

□

#### B.1.4. Simulations studies

##### **Model robustness and comparison**

We performed additional simulations from a Gaussian copula to better understand the sensitivity of the proposed method to choice of copula function. The Gaussian copula was selected as it can capture positive and negative dependence. In the bivariate setting, marginal parameters were set to the same values described in the main text under simulation from a Frank copula. Given that the range of  $\theta$  depends on choice of copula, we aimed to match strength of dependence, in terms of Spearman’s correlation, across copulas rather than magnitude of  $\theta$ . This was done by utilizing the relationship between a copula dependence parameter and Spearman’s correlation. The six  $\theta$  values specified under simulation from a Frank copula  $\{-2.5, -1, 0, 0.5, 1.5, 3\}$  were converted to their equivalent Spearman’s rho and then mapped to the corresponding value of the dependence parameter from a Gaussian copula.

In the multivariate Gaussian setting, the number of features (microbes) was set to 75 to reflect

microbial relative abundance data aggregated to the genus level classification. From these 75 microbes, 2775 pairs can be formed. We assumed that few pairs were truly associated with one another; the off-diagonals of the correlation matrix were generated from a  $\text{Uniform}(0.1, 0.55) \times \text{Bernoulli}(p = 0.015)$ . We further assumed that there is one continuous confounder, drawn from a standard Normal distribution, influencing the zero-inflation probabilities of all microbes. The confounder's corresponding regression coefficient was drawn from a  $\text{Uniform}(1,5)$  distribution. The model intercept was drawn from a  $\text{Uniform}(-1.3,0.5)$  distribution. Correspondingly, the model intercept for mean abundance and dispersion were drawn from  $\text{Uniform}(-4.5,-1.5)$  and  $\text{Uniform}(1,1.5)$  distributions, respectively. The sample size was set as  $n = 100$ .

## B.2. Supplementary figures

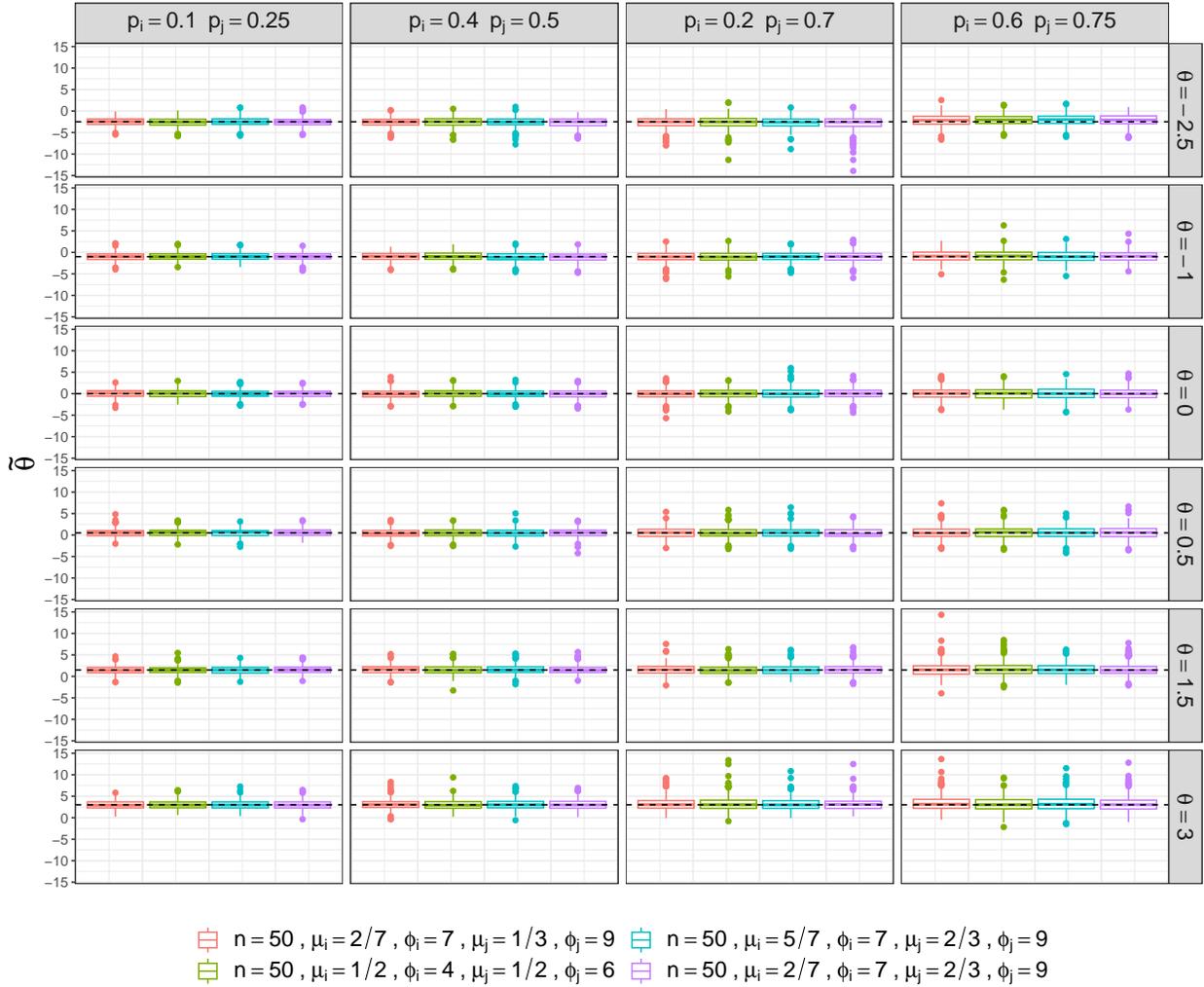


Figure B.1: Boxplots of estimated  $\tilde{\theta}$  values across 500 simulations. The black dashed line represents the true  $\theta$  value. Data was simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ).

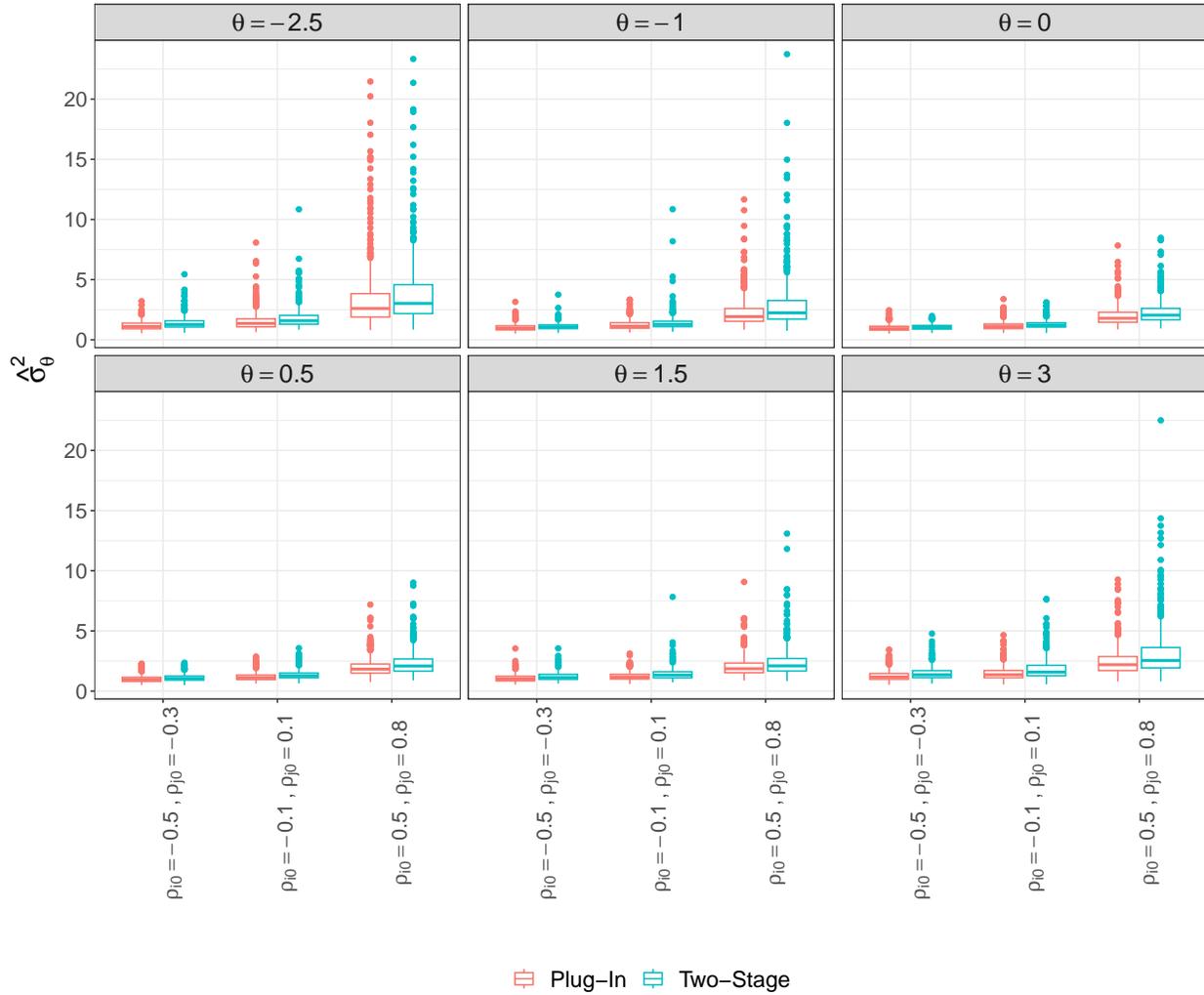


Figure B.2: Boxplots of the jackknife variance of two-stage and plug-in estimated  $\tilde{\theta}$ , denoted as  $\hat{\sigma}_{\tilde{\theta}}^2$ , across 500 simulations. Data was simulated with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ). Outliers with variance values greater than 25 were removed from plots for visualization.

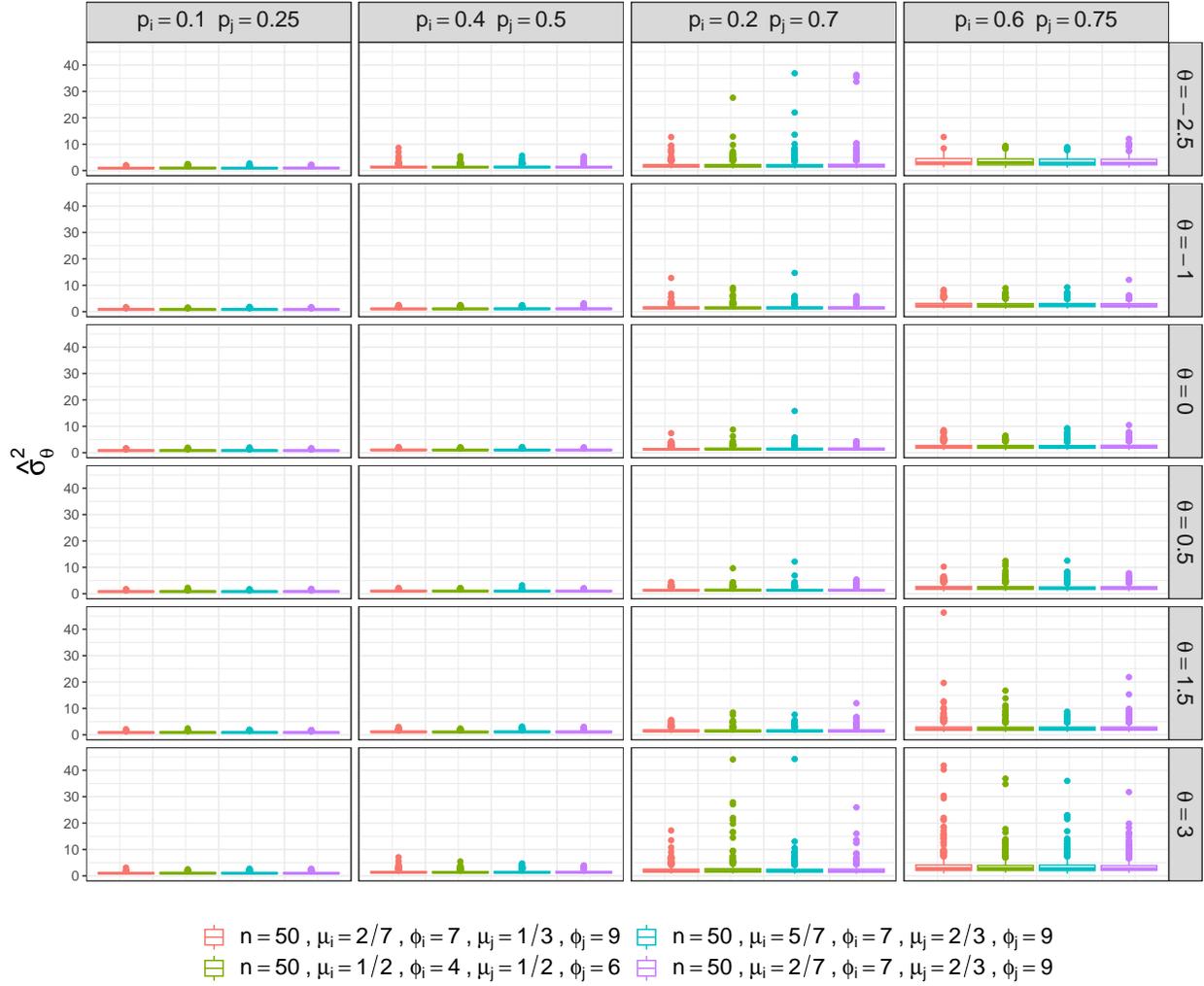


Figure B.3: Boxplots of the estimated jackknife variance of  $\tilde{\theta}$ , denoted as  $\hat{\sigma}_{\tilde{\theta}}^2$ , across 500 simulations. Data was simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ). Outliers with variance values greater than 50 were removed from plots for visualization.

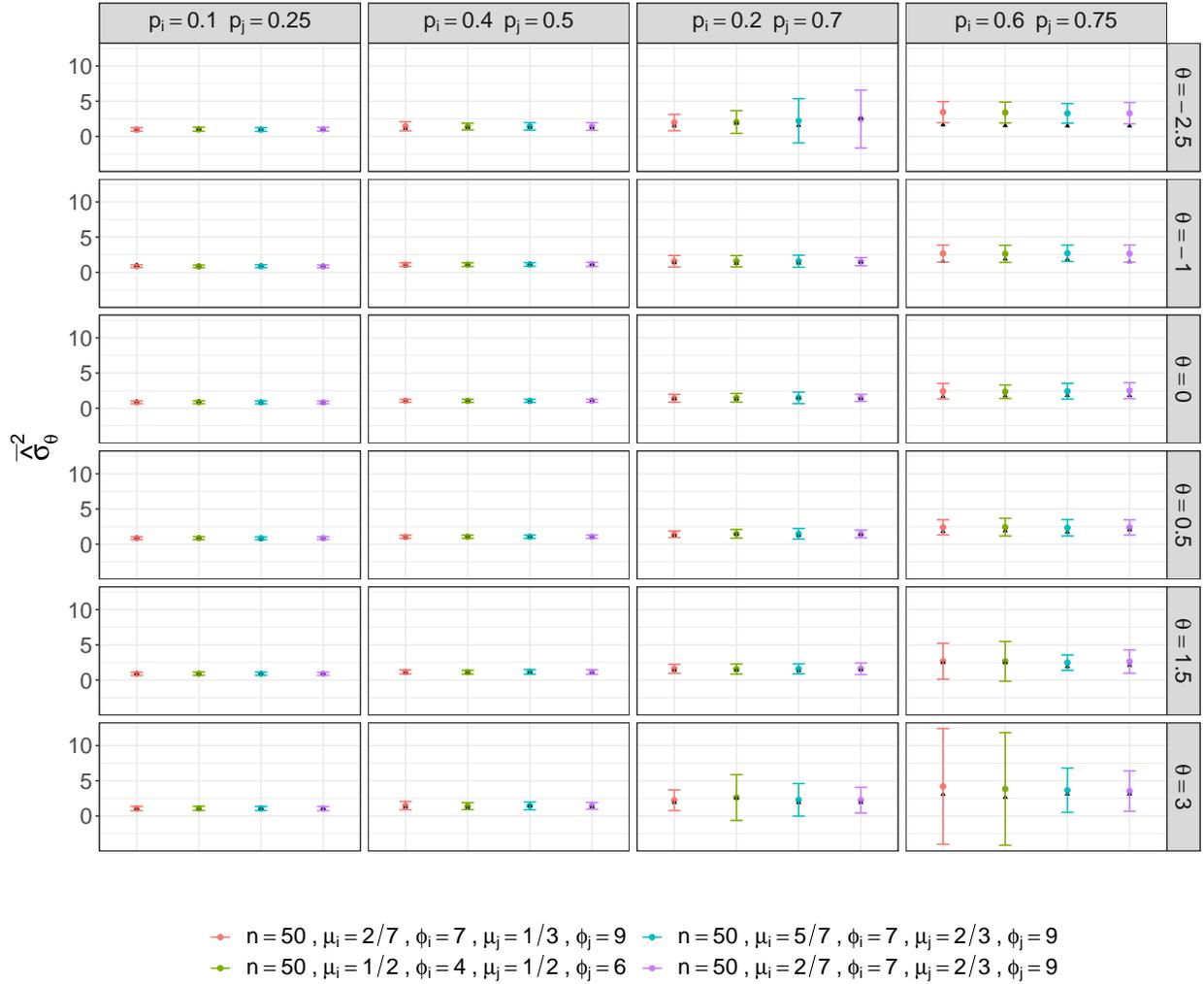


Figure B.4: Mean and standard error bars of the estimated jackknife variance of  $\tilde{\theta}$  from data simulated without covariate adjustment under varying strength of dependence ( $\theta$ ), mean ( $\mu$ ), dispersion ( $\phi$ ) and zero-inflation probability ( $p$ ). Black triangles correspond the empirical (sample) variance.

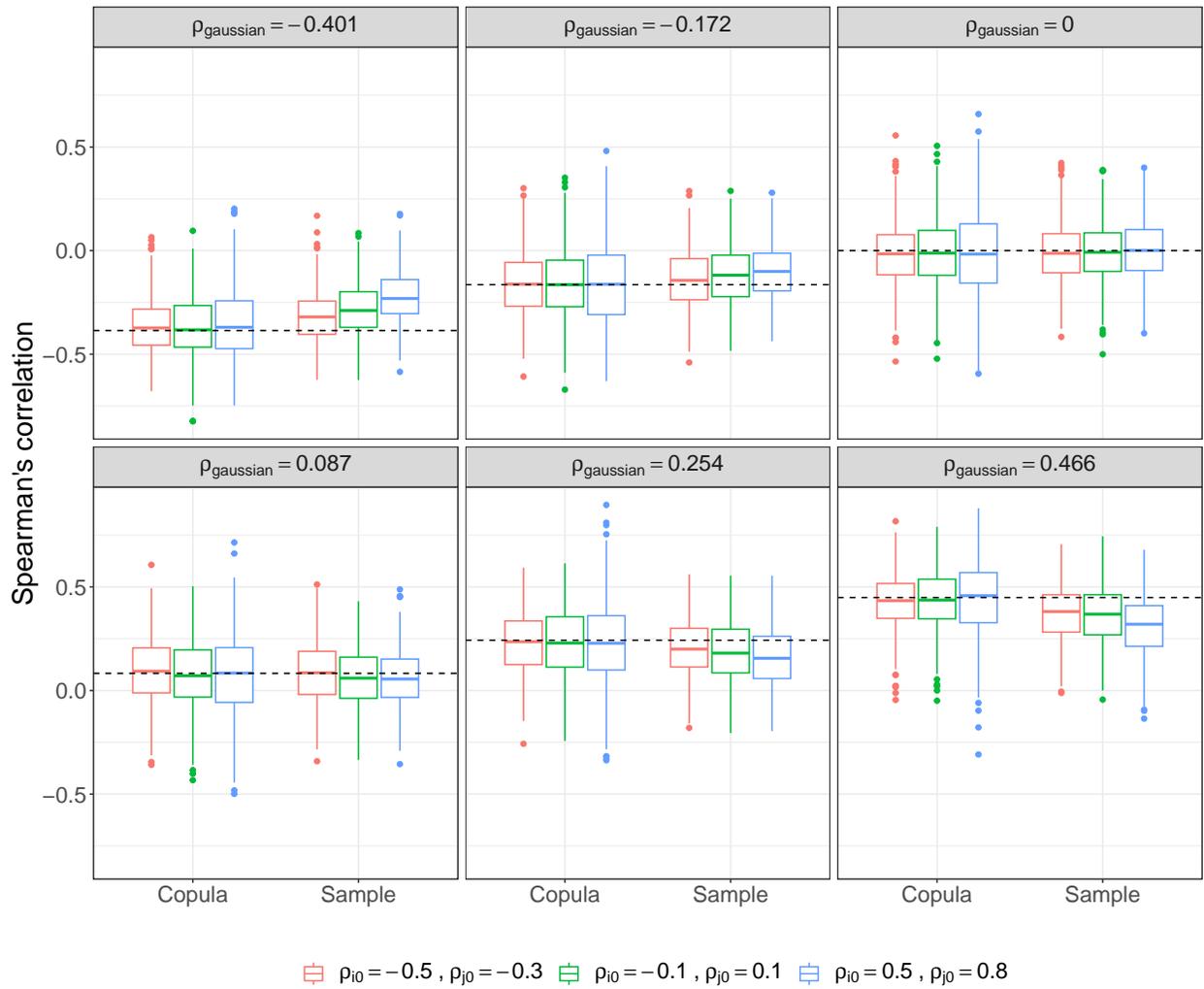


Figure B.5: Boxplot of estimated Spearman's correlation, using copula and sample estimators, across 500 simulations. The black dashed line represents the true value. Data was simulated from a Gaussian copula function with covariate adjustment under varying strength of dependence ( $\theta$ ) and zero-inflation probability ( $\rho_{i0}, \rho_{j0}$ ).

# APPENDIX C

## SUPPLEMENTARY MATERIALS FOR CHAPTER 4

### C.1. Supplementary figures

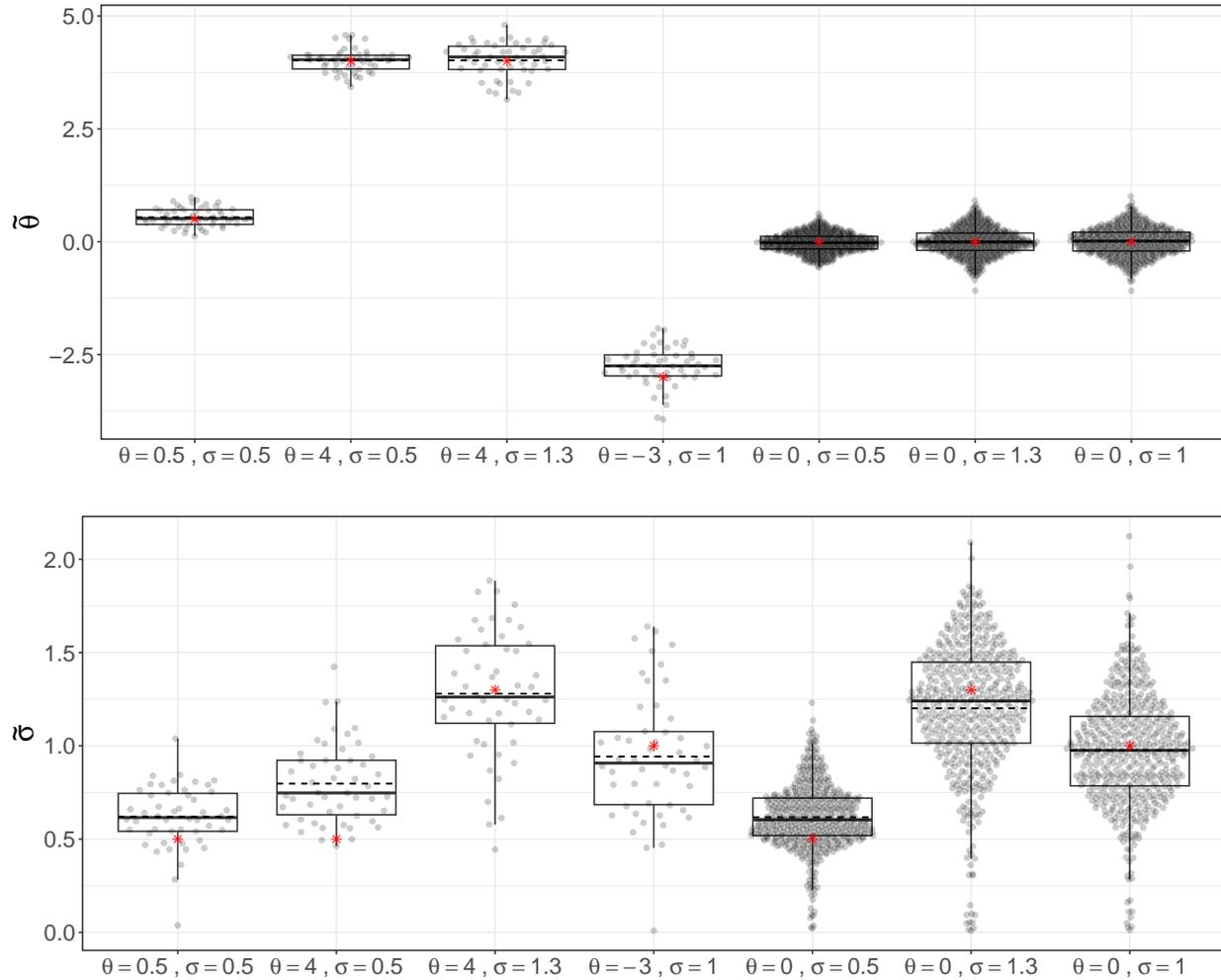


Figure C.1: Boxplots of the  $\tilde{\theta}$  and  $\tilde{\sigma}$  estimates under seven different parameter ( $\theta$  and  $\sigma$ ) settings in simulation and  $n = 50$ . The first four are under the alternative hypothesis ( $\theta \neq 0$  for the Frank copula) and the last three are under the null of no conserved covariation structure ( $\theta = 0$ ). The black dashed line represents the mean across all runs. The red star indicates the true value specified under simulation.

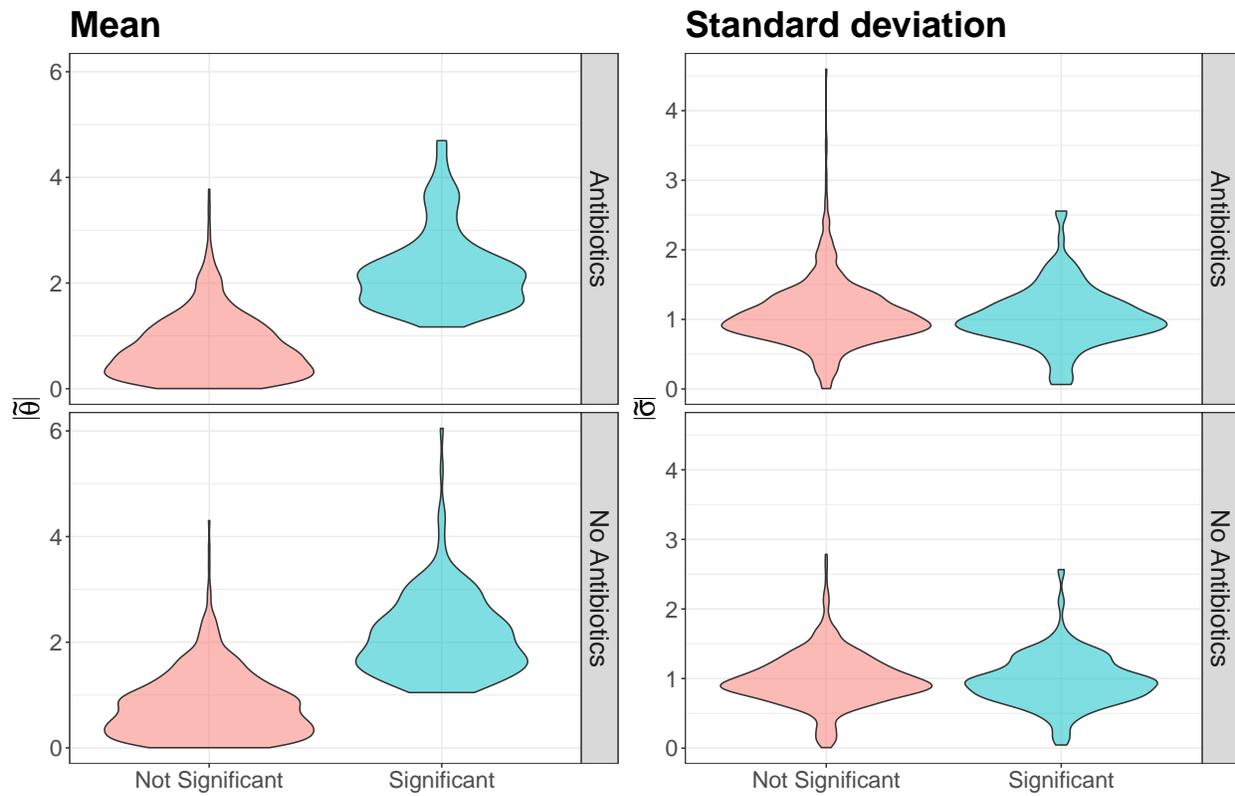


Figure C.2: Violin plots of estimated  $\tilde{\theta}$  and  $\tilde{\sigma}$  from the random-effect for all pairs in the DIABIMMUNE data. Data is split by antibiotics exposure status and significance of the Monte Carlo likelihood ratio test. The distribution of the estimated mean ( $\tilde{\theta}$ ) is shift towards the null value of zero for non-significant pairs. The distribution of the estimated standard deviation ( $\tilde{\sigma}$ ) is similar across significance levels. These trends are consistent across antibiotics exposure status.

## BIBLIOGRAPHY

- J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x.
- Albert Barberán, Scott T. Bates, Emilio O. Casamayor, and Noah Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, 6(2):343–351, 2012. doi: 10.1038/ismej.2011.119.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.
- Heidy M. W. den Besten, Aarathi Arvind, Heidi M. S. Gaballo, Roy Moezelaar, Marcel H. Zwietering, and Tjakko Abee. Short- and long-term biomarkers for bacterial robustness: a framework for quantifying correlations between cellular indicators and adaptive behavior. *PLOS ONE*, 5(10): e13746, 2010. doi: 10.1371/journal.pone.0013746.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. doi: 10.5555/944919.944937.
- M. N. Burge. *Fungi in biological control systems*. Manchester University Press, 1988.
- Yuanpei Cao, Wei Lin, and Hongzhe Li. Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association*, 114(526): 759–772, 2019. doi: 10.1080/01621459.2018.1442340.
- Eric Z. Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 2016. doi: 10.1093/bioinformatics/btw308.
- Jun Chen, Frederic D. Bushman, James D. Lewis, Gary D. Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013. doi: 10.1093/biostatistics/kxs038.
- Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004. doi: 10.1103/PhysRevE.70.066111.
- Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969. doi: 10.2307/2283728.
- Rebecca A. Deek and Hongzhe Li. A zero-inflated latent dirichlet allocation model for microbiome studies. *Frontiers in Genetics*, 11:602594, 2020. doi: 10.3389/fgene.2020.602594.

- Rebecca A. Deek and Hongzhe Li. Inference of microbial interactions using copula models with mixture margins. *arXiv:2111.02344 [stat]*, 2021.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
- Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012. doi: 10.1038/nrmicro2832.
- Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLOS Computational Biology*, 8(7):e1002606, 2012. doi: 10.1371/journal.pcbi.1002606.
- Noah Fierer and Robert B. Jackson. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, 103(3):626–631, 2006. doi: 10.1073/pnas.0507535103.
- Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):e1002687, 2012. doi: 10.1371/journal.pcbi.1002687.
- Marie-Anne Félix and Michalis Barkoulas. Pervasive robustness in biological systems. *Nature Reviews Genetics*, 16(8):483–496, 2015. doi: 10.1038/nrg3949.
- Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.
- Christian Genest, Kilani Ghoudi, and Louis-Paul Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995. doi: 10.1093/biomet/82.3.543.
- Georg K. Gerber. The dynamic microbiome. *FEBS Letters*, 588(22):4131–4139, 2014. doi: 10.1016/j.febslet.2014.02.037.
- Jack A. Gilbert, Janet K. Jansson, and Rob Knight. The Earth Microbiome Project: successes and aspirations. *BMC Biology*, 12(1):69, 2014. doi: 10.1186/s12915-014-0069-1.
- Sharon Greenblum, Peter J. Turnbaugh, and Elhanan Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–599, 2012. doi: 10.1073/pnas.1116053109.
- Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. doi: 10.1073/pnas.030775210.

- David Gunawan, Mohamad A. Khaled, and Robert Kohn. Mixed marginal copula modeling. *Journal of Business & Economic Statistics*, 38(1):137–147, 2020. doi: 10.1080/07350015.2018.1469998.
- Stijn Hawinkel, Frederiek-Maarten Kerckhof, Luc Bijmens, and Olivier Thas. A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, 14(2):e0205474, 2019. doi: 10.1371/journal.pone.0205474.
- Jinzhi He, Yan Li, Yangpei Cao, Jin Xue, and Xuedong Zhou. The oral microbiome diversity and its relation to human diseases. *Folia Microbiologica*, 60(1):69–80, 2015. doi: 10.1007/s12223-014-0342-2.
- Koichi Higashi, Shinya Suzuki, Shin Kurosawa, Hiroshi Mori, and Ken Kurokawa. Latent environment allocation of microbial community data. *PLOS Computational Biology*, 14(6):e1006143, 2018. doi: 10.1371/journal.pcbi.1006143.
- Nhan Thi Ho, Fan Li, Shuang Wang, and Louise Kuhn. metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. *BMC Bioinformatics*, 20(1):188, 2019. doi: 10.1186/s12859-019-2744-2.
- Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLOS ONE*, 7(2):e30126, 2012. doi: 10.1371/journal.pone.0030126.
- Shion Hosoda, Suguru Nishijima, Tsukasa Fukunaga, Masahira Hattori, and Michiaki Hamada. Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. *Microbiome*, 8(1):95, 2020. doi: 10.1186/s40168-020-00864-3.
- Harry Joe. *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, 1997. ISBN 978-0-412-07331-1.
- Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005. doi: 10.1016/j.jmva.2004.06.003.
- Harry Joe and James Jianmeng Xu. The estimation method of inference functions for margins for multivariate models. 1996. doi: 10.14288/1.0225985.
- Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004. doi: 10.1038/nrg1471.
- Hiroaki Kitano. Towards a theory of biological robustness. *Molecular Systems Biology*, 3(1):137, 2007. doi: 10.1038/msb4100179.
- Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and compositionally robust inference of microbial ecological net-

- works. *PLOS Computational Biology*, 11(5):e1004226, 2015. doi: 10.1371/journal.pcbi.1004226.
- Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling interactions in the microbiome: a network perspective. *Trends in Microbiology*, 25(3):217–228, 2017. doi: 10.1016/j.tim.2016.11.008.
- Joshua Lederberg and Alexa T. Mccray. ‘Ome Sweet ‘Omics—a genealogical treasury of words. *The Scientist*, 15(7):8–8, 2001.
- Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015. doi: 10.1146/annurev-statistics-010814-020351.
- Kung-Yee Liang and Steven G. Self. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):785–796, 1996. doi: <https://doi.org/10.1111/j.2517-6161.1996.tb02116.x>.
- William Z. Lidicker, Jr. A clarification of interactions in ecological systems. *BioScience*, 29(8):475–477, 1979. doi: 10.2307/1307540.
- Catherine A. Lozupone and Rob Knight. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences*, 104(27):11436–11440, 2007. doi: 10.1073/pnas.0611525104.
- Daniel McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, Lindsay DeRight Goldasich, Pieter C. Dorrestein, Robert R. Dunn, Ashkaan K. Fahimipour, James Gaffney, Jack A. Gilbert, Grant Gogul, Jessica L. Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A. Jackson, Stefan Janssen, Dilip V. Jeste, Lingjing Jiang, Scott T. Kelley, Dan Knights, Tomasz Kosciolk, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V. Melnik, Jessica L. Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T. Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S. Pollard, Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D. Swafford, Varykina G. Thackray, Luke R. Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanc, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, The American Gut Consortium, and Rob Knight. American Gut: an open platform for citizen science microbiome research. *mSystems*, 3(3), 2018. doi: 10.1128/mSystems.00031-18.
- Lisbeth Olsson, Peter Rugbjerg, Luca Torello Pianale, and Cecilia Trivellin. Robustness: linking strain design to viable bioprocesses. *Trends in Biotechnology*, 40(8):918–931, 2022. doi: 10.1016/j.tibtech.2022.01.004.
- Raydonal Ospina and Silvia L. P. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012. doi: 10.1016/j.csda.2011.10.005.

- Joseph N. Paulson, O. Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013. doi: 10.1038/nmeth.2658.
- Xiaoling Peng, Gang Li, and Zhenqiu Liu. Zero-inflated Beta regression for differential abundance analysis with metagenomics data. *Journal of Computational Biology*, 23(2):102–110, 2015. doi: 10.1089/cmb.2015.0157.
- Arjun S. Raman, Jeanette L. Gehrig, Siddarth Venkatesh, Hao-Wei Chang, Matthew C. Hibberd, Sathish Subramanian, Gagandeep Kang, Pascal O. Bessong, Aldo A. M. Lima, Margaret N. Kosek, William A. Petri, Dmitry A. Rodionov, Aleksandr A. Arzamasov, Semen A. Leyn, Andrei L. Osterman, Sayeeda Huq, Ishita Mostafa, Munirul Islam, Mustafa Mahfuz, Rashidul Haque, Tahmeed Ahmed, Michael J. Barratt, and Jeffrey I. Gordon. A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science*, 365(6449), 2019. doi: 10.1126/science.aau4735.
- Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952. doi: 10.1214/aoms/1177729394.
- Kris Sankaran and Susan P. Holmes. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, 2019. doi: 10.1093/biostatistics/kxy018.
- Janice L. Scealy and Alan H. Welsh. Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):351–375, 2011. doi: <https://doi.org/10.1111/j.1467-9868.2010.00766.x>.
- Jose U. Scher and Steven B. Abramson. The microbiome and rheumatoid arthritis. *Nature Reviews Rheumatology*, 7(10):569–578, 2011. doi: 10.1038/nrrheum.2011.121.
- Joanna H. Shih and Thomas A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, 1995. doi: 10.2307/2533269.
- Abe Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959. doi: 10.2139/ssrn.4198458.
- Guilhem Sommeria-Klein, Lucie Zinger, Eric Coissac, Amaia Iribar, Heidi Schimann, Pierre Taberlet, and Jérôme Chave. Latent Dirichlet Allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Molecular Ecology Resources*, 20(2): 371–386, 2020. doi: 10.1111/1755-0998.13109.
- Jörg Stelling, Uwe Sauer, Zoltan Szallasi, Francis J. Doyle, and John Doyle. Robustness of cellular functions. *Cell*, 118(6):675–685, 2004. doi: 10.1016/j.cell.2004.09.008.
- Veena Taneja. Arthritis susceptibility and the gut microbiome. *FEBS Letters*, 588(22):4244–4249, 2014. doi: 10.1016/j.febslet.2014.05.034.

- Zheng-Zheng Tang and Guanhua Chen. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2019. doi: 10.1093/biostatistics/kxy025.
- Elizabeth A. Thompson. Monte Carlo likelihood in genetic mapping. *Statistical Science*, 9(3): 355–366, 1994. doi: 10.1214/ss/1177010381.
- Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolk, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, and Rob Knight. A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017. doi: 10.1038/nature24621.
- Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, 2007. doi: 10.1038/nature06244.
- Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990. doi: 10.1080/01621459.1990.10474930.
- James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, 5(4):e1000352, 2009. doi: 10.1371/journal.pcbi.1000352.
- Stefanie Widder, Katharina Besemer, Gabriel A. Singer, Serena Ceola, Enrico Bertuzzo, Christopher Quince, William T. Sloan, Andrea Rinaldo, and Tom J. Battin. Fluvial network organization imprints on microbial co-occurrence networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(35):12799–12804, 2014. doi: 10.1073/pnas.1411723111.
- Ryan J. Williams, Adina Howe, and Kirsten S. Hofmockel. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in Microbiology*, 5, 2014. doi: 10.3389/fmicb.2014.00358.
- Zhenjiang Xu and Rob Knight. Dietary effects on human gut microbiome diversity. *British Journal of Nutrition*, 113(S1):S1–S5, 2015. doi: 10.1017/S0007114514004127.
- Moran Yassour, Tommi Vatanen, Heli Siljander, Anu-Maaria Hämäläinen, Taina Härkönen, Samppa J. Ryhänen, Eric A. Franzosa, Hera Vlamakis, Curtis Huttenhower, Dirk Gevers, Eric S. Lander, Mikael Knip, on behalf of the DIABIMMUNE Study Group, and Ramnik J. Xavier.

Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine*, 8(343):343ra81–343ra81, 2016. doi: 10.1126/scitranslmed.aad0917.

Grace Yoon, Irina Gaynanova, and Christian L. Müller. Microbial networks in SPRING - semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in Genetics*, 10, 2019. doi: 10.3389/fgene.2019.00516.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. doi: 10.2202/1544-6115.1128.