Distribution Agreement

Signature:

Zhengyi Zhu	Date

STATISTICAL METHODS FOR ANALYZING MICROBIOME DATA

By

Zhengyi Zhu

Doctor of Philosophy

Biostatistics

Yijuan Hu, Ph.D.
Advisor

Glen A. Satten, Ph.D.
Co-advisor

Zhaohui (Steve) Qin, Ph.D.
Committee Member

Hao Wu, Ph.D.
Committee Member

Accepted:

Kimberly J. Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

STATISTICAL METHODS FOR ANALYZING MICROBIOME DATA

By

Zhengyi Zhu
M.S., Emory University, 2020
B.S., Peking University, 2016

Advisors: Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2022

Abstract

STATISTICAL METHODS FOR ANALYZING CORRELATED MICROBIOME DATA

By

Zhengyi Zhu

Data from studies of the microbiome are accumulating at a rapid rate. The relative ease of conducting a census of bacteria by sequencing the 16S rRNA gene has led to many studies that examine the association between microbiome and health states or outcomes. Many microbiome studies have complex design features (e.g. paired, clustered or longitudinal data) or complexities that frequently arise in medical studies (e.g. the presence of confounding covariates). In this dissertation, we propose novel statistical methods for solving three different problems in microbiome studies – testing microbiome association on matched-set data, combining test results on multiple data scales, and estimating variance-covariance matrix for longitudinal data with missing values.

In the first topic, we address the need for statistical methods for analyzing microbiome data comprised of matched sets, to test hypotheses against traits of interest that vary between members of a set. Matched-set data arise frequently in microbiome studies (e.g. pre- and post-treatment samples from a set of individuals, or data from case participants matched to one or more control participants using important confounding variables). Existing methods can not accommodate complex data such as those with unequal sample sizes across sets, confounders varying within sets, and continuous traits of interest. By leveraging PERMANOVA, a commonly used distance-based method for testing hypotheses at the community level, and the linear decomposition model (LDM) that unifies the community-level and OTU-level tests into one framework, we present a new strategy for analyzing matched-set data. We propose to include an indicator variable for each set as covariates, so as to constrain comparisons between samples within a set, and also permute traits within each set, which can account for exchangeable sample correlations. The flexible nature of PERMANOVA and the LDM allows discrete or continuous traits or interactions to be tested, within-set confounders to be adjusted, and unbalanced data to be fully exploited. We design a wide range of simulations to compare our proposed strategy to alternative strategies, including the commonly used one that utilizes restricted permutation only. We also use simulation to explore optimal designs for matched-set studies. We use our method to analyze data from two real studies to illustrate its flexibility for a variety of matched-set microbiome data.

In the second topic, we propose an approach to integrative analysis of different microbiome data scales using the LDM. Previously, LDM was developed for testing hypotheses (both the community level and the individual taxon level) about the microbiome on 3 scales separately - the relative abundance scale, the arcsin-root-transformed relative abundance scale, the presence-absence scale. LDM also offered an omnibus test (LDM-omni) that combined the results of the relative abundance and arcsin-root-transformed relative abundance scale. In some scenarios, we have observed that the presence-absence analysis worked better

than the initial omnibus test. This suggests the need to develop a new omnibus test that combines results from all three data scales. In order for the omnibus global test to use the best scale at each taxon, we propose an omnibus test based on various p-value combination methods to combine the taxon-level LDM p-values into a statistic we could add to the global LDM test, thus offering optimal power across scenarios with different association mechanisms. The omnibus test is available for the wide range of data types and analyses that are supported by LDM.

In the third topic, we tackle the problem of estimating the variance-covariance matrix of the longitudinal measurements at each taxon. A major challenge of analyzing longitudinal measurements is induced by incomplete data. Incomplete data is a result of missing measurements, e.g., patients are followed for a period of time but miss some of the visits. In such cases, empirical estimation of variance-covariance matrix may not be positive-definite, which is a key feature of a variance-covariance matrix. Thus, there is a need for statistical methods for longitudinal data with missing values, to estimate positive-definite variance-covariance matrix, that accommodate non-normal data distributions, complex missingness patterns and possible constraints on the data (e.g., centered measurements that sum to 0). We develop an algorithm based on a non-parametric model that iteratively optimizes variance-covariance matrix estimation towards the empirical one while parameterizes it in a way such that our variance-covariance matrix estimation is always positive semi-definite. We use simulations and data from a real longitudinal microbiome study to illustrate that our proposed algorithm is robust in a wide range of scenarios.

STATISTICAL METHODS FOR ANALYZING MICROBIOME DATA

By

Zhengyi Zhu
M.S., Emory University, 2020
B.S., Peking University, 2016

Advisors: Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2022

Acknowledgement

Thank you, Dr.Yijuan Hu, for being my PhD advisor. You have taught me enormous lessons not only in doing research, but also in living life. You are my role modal, who has inspired me to work hard and ponder what to pursue in life. It has been the luckiest thing in my whole life to have chosen you as my PhD advisor. Words can not express my gratitude to you.

Thank you, Dr.Glen Satten, for being on my committee and also contributing countless wisdom to my research projects. Your lifelong passion for statistics has deeply impressed me. I really appreciate your patience with me throughout the last six years.

Thank you, Dr.Steve Qin and Dr.Hao Wu, for being my committee members. I have learned a lot from the classes you taught. It has always been a pleasure to talk to you about whatever questions I had. You have been great mentors.

Thank you, Dr.Colleen Kelley, for providing another source of funding. I admire your expertise and dedication to clinical microbiome studies. It has been a pleasure to work with you.

Thank you, all the professors in Emory BIOS department. You are the pillar of our department. You have helped build such a great reputation for Emory BIOS that will always make me proud of being an alumnus.

Thank you, Angela Guinyard, Melissa Sherrer, Mary Abosi, Bob Waggoner, for making Emory BIOS department a better place. You have created a home for all the faculties and students. You have helped me so much on my PhD journey.

Thank you, all the friends I have made in Emory BIOS departments, especially Yingtian Hu, Yunxiao Li, Denis Whelan, Jin Ming. You have made my graduate school life a joyful ride. I will always cherish our friendship.

Thank you, all the friends I have met at Georgia Tech Catholic Center, especially my sponsor Liam Byrne and Fr.Branson Hipp. You have showed me how to live a religious life in the secular world. Your have brought me closer to God.

Thank you, mom and dad, for your unconditional love and support. You have sacrificed so much to provide me the best education. I will try my best to provide you a happy retired life.

Thank you, God, for sending me all these beautiful souls. Thank you for everything I have had in life. Thank you for being with me in good times and in bad.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of microbiome data

Microbiome data is generated through 16S rRNA gene sequencing and shotgun metagenomic sequencing. The bioinformatics tools include the pipeline QIIME and mothur. For example, after preprocessing the raw sequences, two ways are available to generate analyzable microbiome data. The 16S sequences are either mapped to an existing phylogenetic tree in a taxonomydependent way or clustered into OTUs (operational taxonomic units) according to similarity in a taxonomic-independent way. The first way uses the existing phylogenetic tree structure to generate microbiome datasets, whereas the second way clusters sequence reads based on similarity level and then assigns them to different taxonomic levels. In the second way, the reads from the amplicons are clustered into OTUs, based on sequence similarity, and then OTUs are hierarchically assigned to a taxonomic tree at the kingdom, phylum, class, order, family, genus and species ranks using available methods for accurate taxonomy assignments, including BLAST (Altschul et al., 1990), the online Greengenes (DeSantis et al., 2006) and RDP (Cole et al., 2003) classifiers, and phylogenetic tree-based and multimer clustering tree-based methods. The final data produced by taxonomy assignments are tables of read counts (bacterial taxa) that are assigned to nodes of a known taxonomic tree. The tables of read counts or relative abundance quantified from the read counts can be used for analyzing and modeling the microbiome composition (Yinglin Xia, 2018).

1.2 Statistical analysis of microbiome data

There are mainly two themes in the current microbiome studies: (1) to characterize the relationship between microbiome features and biological, genetic, clinical or experimental conditions; and (2) to identify potential biological and environmental factors that are associated with microbiome composition. The goal of these studies is to understand mechanisms of host genetic and environmental factors that shape microbiome. Insights gained from the studies potentially contribute to the development of therapeutic strategies in modulating the microbiome composition in human diseases (Yinglin Xia, 2018).

Microbiome communities in an environmental context can be analyzed by multivariate statistical methods or models. Many statistical models and methods are available for analyzing the association of microbiome community composition and environmental covariates and outcomes. Statistical methods for analyzing microbiome data seem to fall into one of two camps. One camp comprises methods that test the global effect of the microbiome, such as PERMANOVA (McArdle and Anderson, 2001), MiRKAT (Zhao et al., 2015), aMiSPU (Wu et al., 2016) and pairNM (Shi and Li, 2017), which can be used to test the hypothesis that variables of interest (e.g. case–control status) are significantly associated with overall microbial compositions. However, these methods do not provide convenient tests of the effects or contributions of individual operational taxonomic units (OTUs), should a global microbiome effect be found (here we use 'OTU' generically to refer to any feature such as amplicon sequence variants or any other taxonomic or functional grouping of bacterial sequences). The other camp is comprised of OTU-by-OTU tests, often directly using a method developed for RNA-Seq data such as DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) or a modification thereof such as MetagenomeSeq (Paulson et al., 2013); some other methods in this camp have adopted a compositional data approach [such as ANCOM (Kaul et al., 2017; Mandal et al., 2015) and ALDEx2 (Fernandes et al., 2014)] were developed for longitudinal data [such as ZIBR (Chen and Li, 2016)], or employed a multi-stage strategy [such as massMap (Hu et al., 2018)]. While some of these approaches have been widely applied, they generally do not give a single test of the global null hypothesis. Although test statistics or Pvalues from OTU-specific tests can of course be combined to give a global test, the performance of this kind of global test is often poor since many of the OTU-specific tests only contribute noise. Hu and Satten (2020) developed the linear decomposition model (LDM) for analyzing microbial count or relative abundance data that are obtained in a 16S rRNA study or a shotgun metagenomics sequencing study. The LDM gives a unified approach that allows both global testing of the overall effect of the microbiome on arbitrary traits of interest, while also providing OTU-specific tests that correspond to the contribution of individual OTUs to the global test results. It allows for complex fixed-effects models such as models that include multiple variables of interest (both continuous and categorical), their interactions, as well as confounding covariates. It is permutation-based, and so can accommodate clustered data and maintain validity for small sample sizes and when data are subject to overdispersion. Because the permutations are based on the Freedman–Lane approach (Freedman and Lane, 1983), powerful type III or 'last variable added' tests like those used in most linear regression packages (Kleinbaum et al., 2007; Muller and Fetterman, 2012) can be constructed.

1.3 Integrative analysis of microbiome data

Although microbiome data usually comes in tables of read counts, transformation or normalization needs to be performed before analysis due to differences in read depths. It is generally accepted that the analysis based on relative abundance data will work best when associated taxa are abundant, while the analysis based on presence-absence data will work best when associated taxa are less abundant. Because the association mechanism is not known a priori, one strategy is to conduct analysis at each taxon scale separately and combine the results in an omnibus test. For example, PERMANOVA-S (Tang, Chen and Alekseyenko, 2016) and MiRKAT-O (Zhao et al., 2015) provided omnibus tests that combine results from analyzing multiple distance matrices, such as the weighted and unweighted UniFrac distances in a phylogenetic-tree-based approach or Bray-Curtis and Jaccard distances in a non-tree-based approach; note that the weighted UniFrac and Bray-Curtis distances are based on relative abundance data while the unweighted UniFrac and Jaccard distances are based on presence-absence data. LDM was initially developed for taxon data at the relative abundance scale and the arcsin-root-transformed relative abundance scale (which is variance-stabilizing for Multinomial and Dirichlet-Multinomial count data), and also offered an omnibus test that combined the results of the two taxon scales (Hu and Satten, 2020). The LDM applied to the untransformed data works better when associated taxa are abundant and the LDM applied to the transformed data works better when associated taxa are less abundant. More recently, LDM was extended for analyzing data at the presence-absence scale (Hu, Lane and Satten, 2021), which accounted for variability of library size by a rarefaction-based yet non-stochastic approach that evaluated the expected LDM test statistic over all rarefaction replicates. The presence-absence analysis performs better than the relative-abundance-based analysis when associated taxa are more rare. Therefore, we might gain extra power by combining results from all three taxon data scales, which has not been explored before.

1.4 Matched-set microbiome data

Many studies of the microbiome have matched-pair or matched-set designs, in which data naturally cluster into sets but the samples within each set vary in the traits of interest (e.g., clinical outcomes, environmental factors). Matching allows us to study the association between the microbiome and the traits of interest by comparing samples *within* sets, ignoring the variability in microbiomes *between* sets. For example, we may collect paired samples pre and post treatment from a set of subjects to assess the treatment effects on the microbiome. We may also collect matched case-control subjects who were matched on important confounding factors to facilitate case-control comparison. Matching is advantageous when the signal-to-noise ratio is larger within than between sets. In matched studies, complexities may occur when the data are *unbalanced* (e.g., having unequal ratio of case-to-control samples per set), there exist additional confounders that vary within each set, or some traits of interest are continuous.

Only two methods have been developed specifically for analyzing matched-set microbiome data; both are limited to paired data without any within-pair covariates. Shi and Li (2017) proposed a paired-multinomial distribution that is only applicable when the sample size is larger than the number of taxa. Zhao et al. (2018) developed a generalized paired Hotelling's test that relaxed the restriction of Shi and Li's method, but can only provide tests at the community level. Matched-set data may be also be considered as a special case of longitudinal data with an exchangeable correlation; as a result, some methods for analyzing longitudinal data can be used to analyze matched-sets microbiome data. These methods are applied separately to each OTU (operational taxonomic unit; here we use "OTU" generically to refer to any feature such as amplicon sequence variants or taxonomic/functional grouping of microbial sequences). An appealing choice is the linear mixed-effects model (LMM), which has typically been applied to arcsin-root-transformed relative abundance data to improve normality (La Rosa et al., 2014; Bokulich et al., 2018; Vatanen et al., 2018). A zero-inflated Beta regression model with random effects (ZIBR) has also been developed specifically for modelling (untransformed) relative abundance data (Chen and Li, 2016). Both methods are based on fully parametric models and so may not fit every OTU well. Some methods have also been developed specifically for analyzing matched-set microbiome data but are limited to paired data without any within-pair covariates. Shi and Li (2017) proposed a paired-multinomial distribution that is only applicable when the sample size is larger than the number of taxa. Zhao et al. (2018) developed a generalized paired Hotelling's test that relaxed the restriction of Shi and Li's method, but can only provide tests at the community level. Further, some strategies have been proposed to extend existing tests of the microbiome to analyzing matched-set data. DESeq2 (Love et al., 2014), originally a method for RNA-Seq data, has frequently been used for one-OTU-at-a-time analyses of microbiome data. The manual for the DESeq2 software package recommends that indicators of set membership should be included as terms in the design formula, but DESeq2 does not account for within-set correlations. PERMANOVA (McArdle and Anderson, 2001) is a commonly used distance-based method for testing hypotheses at the community level. The documentation for two implementations of PERMANOVA, `adonis2` (R package `vegan`) and `permanovaFL` (R package `ldm` (Hu and Satten, 2020)) that differ in their permutation schemes, suggest that restricted permutation within each set should be performed when analyzing matched-set data. However, the performance of any of these strategies have not yet been evaluated, especially in studies with unbalanced data or within-set confounders. Like PERMANOVA, LDM (Hu and Satten, 2020) is also regression- and permutation-based, making it readily extendable to analyzing matched-set data while accounting for the aforementioned data complexities. Although only within-cluster permutation was considered in the LDM paper (Hu and Satten, 2020), that was in a context in which variables of interest could be below the cluster level. The matched set data was not studied from either a theoretical or numerical point of view.

1.5 Longitudinal microbiome data

The microbiome is inherently dynamic, driven by interactions with the host and the environment, and varies over time. Thus, longitudinal microbiome data analysis provides rich information on the profile of microbiome with host and environment interactions. The distinguishing feature of longitudinal studies is that the subjects are measured repeatedly during the study, allowing the direct assessment of changes in response variable over time. Longitudinal study also captures between-individual differences (heterogeneity among individuals) and within subject dynamics. It offers the opportunity to study complex biological, psychological, and behavioral hypotheses, especially those involving changes over time. The advantage of longitudinal analysis is also suitable for microbiome data. It will enhance our understanding of short-and long-term trends of microbiome by intervention, such as diet, and the development and persistence of chronic diseases caused by microbiome (Yinglin Xia, 2018).

The variance-covariance matrix is an essential component of many algorithms in longitudinal studies. Variance-covariance matrices are real, symmetric positive semi-definite matrices. They arise in situations where covariance between pairs of random variables are computed, and also when pairwise interaction between objects are formed, for example, in longitudinal studies. Estimation of covariance matrices involves the design and analysis of statistical procedures for recovering the covariance matrix from data samples. It is common, in practice, to be faced with an approximate variance-covariance matrix: a matrix that is supposed to be a variance-covariance matrix but for a variety of possible reasons is not. In longitudinal microbiome studies, for example, the covariance may be between samples measured over a period of time for the same patients and missing data (perhaps due to patients not being available for some of the visits) may compromise the covariance and lead to a non-positive semi-definite matrix. Again in longitudinal microbiome studies, the sample size can be so small that there are less subjects with full observations than the number of visits, which again can destroy the positive-semi-definiteness of the variance-covariance matrix. The use of empirical variance-covariance estimation in these applications can render the methodology invalid and lead to negative variances being computed. The prevalence and inevitability of missing data, along with the importance of covariance matrices in many applications and algorithms, makes the study of covariance estimation with missing observations of great importance.

Even though missing data is pervasive, and covariance matrices are utilized by a plethora of algorithms, studies on covariance estimation with missing data are few and of a narrower scope than those considering the complete data case. Perhaps because there are many available estimators, and each of them is designed and analyzed for a different type of missing data mechanism. Schafer (1997) developed an EM algorithm for incomplete data that assumes multivariate normal distribution. Higham (2002) developed an alternating projections method that computes the nearest correlation matrix of a symmetric matrix based on the weighted Frobenius norms, but is more suitable for financial data where a long sequence of correlated data (e.g. stock price) is presented, compared to microbiome longitudinal studies where there are usually less than 10 measurements per subject. Among existing methods, high-dimensionality attracts more attention. Cai and Zhang (2016) proposed a minimax rate-optimal estimation of high-dimensional covariance matrices, where sparse and bandable structures were considered. Lounici (2014) considered both structured (low rank) and unstructured covariance matrices and proposed a simple procedure computationally tractable in high-dimension and that does not require imputation of the missing data. These two works offer error analyses and bounds that match similar ones obtained for complete data. They studied convergence to the population covariance matrix (consistency) of different estimators by deriving finite sample error bounds. They assume that the observations are independent and identically distributed (i.i.d.) copies of a sub-Gaussian vector, while the population covariance matrix may be structured or unstructured. More recently, Park and Lim (2019) considered a non uniform and dependent missing data pattern. This work however, when simplified to the case studied by Lounici (uniform independent observations), returns convergence rates that are sub-optimal with respect to the ambient dimension and the rate of missing entries.

Chapter 2

Constraining PERMANOVA and LDM to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data

2.1 Introduction

In this article, we develop a new strategy for using PERMANOVA and the LDM to analyze a wide range of matched-set microbiome data, for testing both community-level hypotheses and individual OTUs whenever applicable. In the methods section, we describe our strategy and establish a connection with the existing strategy of restricted permutation. In the results section, we present the simulation studies and the application to two real microbiome studies with matched-set designs. We conclude with a discussion section.

2.2 Methods

We will refer to each observation as a "sample" and refer to the experimental unit that contributes one or more observations as a "set". We allow each set to be comprised of an arbitrary number of samples. We also allow multiple discrete and/or continuous traits to be tested and additional sample-level (i.e., within-set) confounding covariates to be adjusted for. In a common scenario with a binary trait (e.g., a case-control status or a treatment or exposure variable), each set consists of one case sample and m ($m \geq 1$) control samples, usually referred to as 1:m matched data. We assume that, after all covariates (including the traits of interest) have been accounted for, the members of each set are exchangeable.

To present our strategy for analyzing matched-set data, we introduce a common notation to describe both PERMANOVA and the LDM. Both PERMANOVA and the LDM are linear models for which the effects of covariates (metadata) are summarized in a design matrix X. The rows of X correspond to samples while the columns of X correspond to the covariates. We may partition X by columns into K groups (which we call "submodels") such that $X = (X_1, X_2, \ldots, X_K)$, where each X_k denotes a variable or set of variables we wish to test (jointly). For example, X_k may consist of indicator variables for levels of a single categorical variable, or a group of potential confounders that we wish to adjust for simultaneously. Both PERMANOVA and the LDM make the columns of X_k *orthonormal* to the columns of $X_{k'}$

for $k' < k$ using projection (i.e., the Gram-Schmidt process). Thus, we require an ordering of the submodels, which leads to unambiguous interpretations of p-values, that is, the test of each submodel is adjusted for the proceding submodels.

For both PERMANOVA and the LDM, test statistics for the kth submodel can be expressed in terms of the quantity $X_k^{\mathrm{T}} Y$. For PERMANOVA, Y is related to the (squared and Gower-centered) distance matrix Δ by $\Delta = YSY^{\mathrm{T}}$, where S is a diagonal matrix with diagonal elements equal to 1 or -1 corresponding to positive and negative eigenvalues of Δ, respectively. For the LDM, Y is the (column-centered) OTU table that has rows for samples and columns for OTUs; the OTU table typically contains the frequency (i.e., relative abundance) data or arcsin-root transformed frequency data. Since Y in either PERMANOVA or the LDM is column-centered and treated as the response of a linear model, we also assume the design matrix X is *column-centered.*

With no loss of generality, we can write the element of Y in the ith row and jth column as

$$Y_{i,j} = \overline{Y}_{s(i),j} + (Y_{i,j} - \overline{Y}_{s(i),j}) = \overline{Y}_{s(i),j} + \delta_{i,j},$$

where $s(i)$ is the set that the ith sample belongs to. Thus $\overline{Y}_{s(i),j}$ is the set-level average of $Y_{i,j}$ and $\delta_{i,j}$ is the deviation of the ith sample from the set-level average. The rationale of a matched-set design is that we wish to treat $\overline{Y}_{s(i),j}$ characterizing a set as a *nuisance parameter* and focus the testing efforts on $\delta_{i,j}$s. With this in mind, we note that $X_k^{\mathrm{T}} Y$ is a function of only the $\delta_{i,j}$s (i.e., not a function of the $\overline{Y}_{s(i),j}$s) whenever the column values of X_k sum to zero for each set of samples belonging to the same set. It is clear that this occurs whenever the columns of X_k are orthogonal to the set of indicator variables corresponding to the set IDs. Therefore, our proposed strategy for fitting matched-set data is to introduce an indicator variable for each set to be included in submodel X_1 along with any sample-level confounding covariates that are not matched on. Note that any set-level confounders are automatically controlled for in this strategy, as they can be written as linear combinations of the indicator variables generated by the set IDs. Indeed, it is typical of matched-set analyses that the effect of variables that have been matched on (i.e., that are constant in each set) cannot be determined (see e.g., (Breslow et al., 1980)).

To see how this works in practice, consider a simple example with two sets, the first having two samples and the second having three samples. For clarity, we work with X_ks before orthonormalization and show X_1 (which has the indicator variables for the two sets) and X_2 (which has a case-control status) before column centering:

$$X_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

After column centering (i.e., subtracting column means), $X_2 = (3/5, -2/5, 3/5, -2/5, -2/5)^{\mathrm{T}}$. Note that the values in X_2 do not sum to zero within each set. If we constructed a test using this contrast, the set-specific means $\overline{Y}_{1,j}$ and $\overline{Y}_{2,j}$ would not be eliminated. However, if we make X_2 (before column centering) orthogonal to the columns of X_1 (which automatically achieves column centering), we find $X_2 = (1/2, -1/2, 2/3, -1/3, -1/3)^{\mathrm{T}}$, where we see that the values in X_2 sum to zero within each set.

We identify a condition under which the nuisance parameters disappear even without projecting off the set ID. We say that a variable in matched-set data is *balanced* if the sum of the variable within each set is proportional to the number of its samples (with the same constant of proportionality). For example, a case-control status is balanced if all sets have as many case as control samples, or if some sets have two case and four control samples and the remaining sets have one case and two control samples. *For a balanced variable, column centering alone is sufficient to make the values of that variable sum to zero within each set, even without projecting off the set ID.* Note that adjusting for sample-level covariates can result in imbalance in a variable, even if it was initially balanced; in this case, projection on the set ID is required to restore balance. A simple example with two sets, each contributing two samples along with a sample-level covariate, shows this. Before column centering (and orthonormalization), suppose the covariate is $X_1 = (9, 8, 6, 9)^{\mathrm{T}}$ and the case-control status is $X_2 = (1, 0, 1, 0)^{\mathrm{T}}$. After column centering we have

$$X_1 = \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1/2 \\ -1/2 \\ 1/2 \\ -1/2 \end{pmatrix}.$$

In the absence of the covariate, X_2 after column centering do sum to zero within each set; however, after adjusting for the covariate we have $X_2 = (2/3, -1/2, 1/6, -1/3)^{\mathrm{T}}$, which does not eliminate the set-specific means. If we also adjust for the set ID by augmenting X_1 with the column-centered indicator

$$X_1 = \begin{pmatrix} 1/2 & 1 \\ 1/2 & 0 \\ -1/2 & -2 \\ -1/2 & 1 \end{pmatrix},$$

we obtain $X_2 = (3/5, -3/5, 1/5, -1/5)^{\mathrm{T}}$, in which values do sum to zero within sets. Finally, note that in this example we have considered a binary case-control trait; it should be clear that, for a continuous trait, the within-set sum is unlikely to be the same for each set, and hence the projection on the set ID is required to eliminate the nuisance parameters.

As we have assumed the samples in each set are exchangeable, we propose to perform restricted permutation among samples from the same set. As permuting residuals of Y in the Freedman and Lane scheme (Freedman and Lane, 1983) is typically equivalent to permuting X_ks (Hu and Satten, 2020), the restricted permutation refers to permuting the (orthonormalized) traits of interest within samples from the same set. The same permutation scheme can be used for testing the interactions between traits of interest or between traits and set-level covariates; the latter allows us to detect whether the associations between the microbiome and the traits of interest are homogeneous across study groups (which can be defined by the set-level covariates). As noted previously, when *all* variables are balanced, the columns of X (excluding the set ID indicator vectors) will automatically be orthogonal to the set ID indicator vectors. Since both `permanovaFL` and the LDM permute the rows of X, it is also clear that this orthogonality holds for every permutation as long as permutations are conducted within sets. As a result, the p-values for `permanovaFL` or the LDM will be identical with and without adjustment of the set ID in this situation, as long as the restricted permutation is performed.

2.3 Results

2.3.1 Simulation studies

To generate our simulation data, we used the same motivating dataset as Hu and Satten (Hu and Satten, 2020), i.e., data on 856 OTUs of the upper-respiratory-tract microbiome first described by Charlson et al. (Charlson et al., 2010). In most simulations we considered a binary trait such as case-control status, but we also considered matched sets with a continuous trait. We defined a "causal" OTU to have frequency that depended on the trait of interest. We considered on two complementary causal mechanisms: the first mechanism (S1) assumed that half (428) of the OTUs (after excluding the three most abundant OTUs) were causal; the second mechanism (S2) assumed the ten most abundant OTUs were causal. In each scenario, we randomly partitioned the causal OTUs into two equal-size subsets, S_{-}^{trait} and S_{+}^{trait}, to contain OTUs with decreased and increased frequencies, respectively, in cases relative to controls. We further partitioned S_{+}^{trait} into S_{+a}^{trait} and S_{+m}^{trait} comprised of OTUs whose frequencies are increased in additive and multiplicative manners, respectively. For the simple situation, with no covariates but the trait of interest, we simulated data for the ith set using the following steps.

1. We assigned trait values X_{ij}^{trait} to the jth sample of the ith set. For matched pair samples, $X_{ij}^{\text{trait}} = 0$ was assigned to control samples and $X_{ij}^{\text{trait}} = 1$ to case samples; for continuous traits, X_{ij}^{trait} was sampled from the $U[0, 1]$ distribution.

2. We generated the mean OTU frequencies $\overline{\pi}_i$ for set i from the Dirichlet distribution $Dir(\overline{\pi}, \theta_1)$, where the mean parameter $\overline{\pi}$ and overdispersion parameter θ_1 took the values of the estimated mean and overdispersion (0.02) in the Dirichlet-Multinomial (DM) model fitted to the upper-respiratory-tract data. Note that $\overline{\pi}$ and θ_1 characterize the population mean of OTU frequencies and between-set heterogeneity.

3. Given $\overline{\pi}_i$, we generated the baseline OTU frequencies $\pi_{ij}^{(0)}$ for sample j of set i from the Dirichlet distribution $Dir(\overline{\pi}_i, \theta_2)$, where θ_2 was set to 0.007, which was the median of the estimated overdispersion in the DM model that was fitted to data for each set with three samples in the MsFLASH study (see the section "Analysis of the MsFLASH data"). Note that θ_2 characterizes heterogeneity among samples from the same set, and $\pi_{ij}^{(0)}$ represents the (true) OTU frequencies we would see when trait $X_{ij}^{\text{trait}} = 0$.

4. We then generated the (true) OTU frequencies that account for a non-zero effect of trait X_{ij}^{trait}, denoted π_{ij}^{trait}, by reducing the frequency of each OTU in S_{-}^{trait} by a factor of β, then distributing half of the total reduced frequency evenly to OTUs in S_{+a}^{trait} and the other half to OTUs in S_{+m}^{trait} in proportion to their baseline frequencies in $\pi_{ij}^{(0)}$. We then formed the (true) OTU frequency for the jth sample from the ith set using $\pi_{ij} = (1 - X_{ij}^{\text{trait}})\pi_{ij}^{(0)} + X_{ij}^{\text{trait}}\pi_{ij}^{\text{trait}}$. Note that β characterizes the *effect size* of the trait, i.e. the amount by which OTU frequencies vary at the causal OTUs when the trait $X_{ij}^{\text{trait}} = 1$.

5. We generated read count data for each sample using the multinomial distribution $MN(\pi_{ij}, N_{ij})$, where the total read count N_{ij} was generated from the Poisson distribution with mean 10000 (and set to 500 if the Poisson sampling resulted in a value less than 500).

To induce the effects of additional covariates, we made further modifications to $\overline{\pi}_i$ and/or π_{ij} that were similar to the modifications made to $\pi_{ij}^{(0)}$ to construct π_{ij}^{trait}. For simulations where we wished to include a main effect of a set-level covariate X_i^{set}, we first sampled values of X_i^{set} from a Bernoulli distribution with parameter 0.5. We then uniformly sampled 428 OTUs to be associated with the covariate, and randomly partitioned them into two equal-size subsets S_-^{set} and S_+^{set}. We then constructed $\overline{\pi}_i^{\text{set}}$ by modifying $\overline{\pi}_i$, reducing the frequency of each OTU in S_-^{set} by a factor of $\beta^{\text{set}} = 0.2$ and distributing the total reduced frequency to OTUs in S_+^{set} in proportion to their original frequencies in $\overline{\pi}_i$. We then replaced $\overline{\pi}_i$ by $(1 - X_i^{\text{set}})\overline{\pi}_i + X_i^{\text{set}}\overline{\pi}_i^{\text{set}}$, to be used in Step 3.

To account for a sample-level confounder X_{ij}^{sam}, we first sampled X_{ij}^{sam} from a Bernoulli distribution with parameter $(0.2 - 0.1X_{ij}^{\text{trait}})$. We then uniformly sampled 428 OTUs to be associated with the covariate, and randomly partitioned them into two equal-size subsets S_-^{sam} and S_+^{sam}. We then constructed π_{ij}^{sam} by modifying π_{ij} in the same way that $\overline{\pi}_i^{\text{set}}$ was modified from $\overline{\pi}_i$, but with a factor of $\beta^{\text{sam}} = 0.5$. We then replaced π_{ij} by $(1 - X_{ij}^{\text{sam}})\pi_{ij} + X_{ij}^{\text{sam}}\pi_{ij}^{\text{sam}}$. The resulting values were used in Step 5.

Finally, to account for an interaction between a set-level covariate and the trait, we sampled a third set of OTUs (a random sample of 428 OTUs under S1 and the top 1-5 and 11-15 most abundant OTUs under S2) to be associated with the interaction, and randomly partitioned them into two equal-size subsets S_-^{int} and S_+^{int}. Then, when both $X_i^{\text{set}} = 1$ and $X_{ij}^{\text{trait}} = 1$, we further modified π_{ij} by reducing the frequency of OTUs in S_-^{int} by a factor β^{int} and then distributing this extra mass to S_+^{int} in proportion to the OTU frequencies in π_{ij}. The resulting values of π_{ij} were then used in Step 5. Note that whenever we included an interaction term like this, the main effect of X_i^{set} ($\beta^{\text{set}} = 0.2$) and X_{ij}^{trait} ($\beta = 0.5$) was also included as described previously.

We evaluated the performance of different strategies and methods in seven scenarios of matched-set data: (1) matched-pair data, (2) unbalanced data, (3) matched-pair data with a sample-level confounder, (4) matched-pair data with a set-level covariate, (5) unbalanced data with a set-level covariate, (6) matched-pair data with a continuous trait, and (7) matched-pair data with an interaction effect. To facilitate comparison across scenarios, the same sets of causal OTUs ($S_-^{\text{trait}}, S_+^{\text{trait}}$) and covariate-associated OTUs ($S_-^{\text{set}}, S_+^{\text{set}}$), ($S_-^{\text{sam}}, S_+^{\text{sam}}$) and ($S_-^{\text{int}}, S_+^{\text{int}}$) (if called for), were used for all scenarios. For each scenario except (2), (5) and (6), we generated data for 50 1:1 matched pairs (with a binary trait); for scenarios (2) and (5) with unbalanced data we generated data for 25 1:1 matched pairs and 25 1:2 matched sets (with a binary trait); for scenario (6), we generated data for 50 matched pairs with a continuous trait.

We also explored various 1:m matched study designs to assess the performance under varying conditions. First, we compared the design that collected 50 1:1 matched pairs with the design that collected 50:50 independent case-control samples (first simulating pairs and then selecting only one sample from each pair), over varying values for the within-set heterogeneity θ_2. Second, we compared different 1:m matched-set designs with a fixed total of 90 samples. Specifically, we considered $m = 1$, 2, 4, and 5 and collected 45 1:1 pairs, 30 1:2 sets, 18 1:4 sets, and 15 1:5 sets, respectively, to form each dataset. We also considered $m = 3$ and collected 22 of 1:3 sets and 1 pair (to meet the total sample size 90) for the 1:3 design. Lastly, we compared different 1:m ($m = 1, 2, 3, 4, 5$) designs when fixing the total number of sets to 50.

We applied PERMANOVA (implemented in both `permanovaFL` and `adonis2`) and the LDM with the proposed strategy (*adjusting for the set ID and sample-level covariates if present, not adjusting for set-level covariates, and performing restricted permutation within sets*). PERMANOVA tests were calculated using the Bray-Curtis distance unless otherwise noted. We report LDM results for the omnibus test that combines the test results from raw frequency (relative abundance) data and arcsin-root-transformed frequency data (Hu and Satten, 2020). For testing individual OTUs, we compared the LDM with the proposed strategy to the following alternative methods: LDM without adjusting for the set ID; LDM without performing restricted permutation; DESeq2 (adjusting for set ID); LMM (applied to arcsin-root-transformed relative abundance data); ZIBR when it is applicable (i.e., for data with equal number of samples in each set); and the Wilcoxon signed-rank test when it is applicable (i.e., for matched-pair data). We evaluated the type I error and power for the community-level (global) test of any microbiome effect at nominal significance level 0.05, and we assessed empirical sensitivity (proportion of truly associated OTUs that were detected) and empirical FDR for the OTU tests at a nominal FDR of 10%. Results for type I error were based on 10000 replicates; all other results were based on 1000 replicates. OTUs having fewer than 5 non-zero entries were removed before analysis.

2.3.2 Simulation results

Results on type I error for the seven scenarios we considered were summarized in Table 2.1. The results of power, sensitivity, and FDR for the seven scenarios were displayed in Figures 2.1–2.7, respectively. In all scenarios, our proposed strategy, when applied to either `permanovaFL` or the LDM, yielded correct type I error and the highest power compared to alternative strategies; `adonis2` with the proposed strategy produced slightly conservative type I error and slightly lower power compared to `permanovaFL`. The LDM using the proposed strategy always controlled the FDR and achieved the highest sensitivity compared to the LDM using alternative strategies or DESeq2 when it controlled the FDR or Wilcoxon if it is applicable. The ZIBR method always yielded highly inflated FDRs. With a binary trait, the LMM always resulted in conservative FDR and diminished sensitivity compared to the LDM with the proposed strategy; with a continuous trait, conversely, it led to inflated FDRs.

For (1) the matched-pair data, `permanovaFL` and the LDM not adjusting for the set ID produced identical results to their counterparts using the proposed strategy as expected. Note that p-values from `adonis2` were not identical with and without adjustment for set ID, but the type I error and power (Figure 2.1) of the two strategies were very similar. Here and for all datasets with a binary case-control trait, the strategy of performing *unrestricted* permutation led to conservative type I error and FDR and diminished power and sensitivity

(Figures 2.1–2.5, 2.7) when applied to `permanovaFL` and the LDM, but inflated type I error when applied to `adonis2`.

For (2) the unbalanced data, the LDM not adjusting for the set ID yielded correct type I error but diminished power and sensitivity relative to its counterpart that using the proposed strategy (Figure 2.2). The same pattern can be seen in the results of `permanovaFL` and `adonis2`.

For (3) the matched-pair data with a sample-level confounder, `permanovaFL`, `adonis2`, and the LDM not adjusting for the confounder had inflated type I error ($0.063 \sim 0.080$), indicating that we have indeed induced some confounding effect in the data. In the presence of such a confounding effect, `permanovaFL` and the LDM not adjusting for the set ID (even after adjusting for the confounder) had inflated type I error ($0.071 \sim 0.087$). In this case, not adjusting for the set ID did not just affect the power, but also affected the validity. These methods with inflated type I error were not included in Figure 2.3. In contrast, `adonis2` not adjusting for the set ID had more conservative type I error than that adjusting for the set ID (both adjusting for the confounder), so the former had reduced power compared to the latter.

Our proposed strategy was robust to the presence of set-level covariates. For (4) the matched-pair data with a set-level covariate, whether or not adjusting for the covariate or the set ID (i.e., the first three strategies in scenario (4) of Table 2.1) all yielded identical results when applied to `permanovaFL` or the LDM, as we have analytically shown. Thus Figure 2.4 only displayed their results for the proposed strategy. When applied to `adonis2`, the three strategies led to slightly different type I error and power. For (5) the unbalanced data with a set-level covariate, the LDM not adjusting for the set ID generated correct type I error but diminished power and sensitivity compared to its counterpart that adjusted for the set ID (both not adjusting for the covariate) (Figure 2.5). Adjusting for the covariate but not the set ID failed to recover any power or sensitivity, which understored the importance of adjusting for the set ID. The same pattern can be seen in the results of `permanovaFL` and `adonis2`.

In the presence of a continuous trait, even in (6) the simplest matched-pair data without any covariates, `permanovaFL`, `adonis2`, and the LDM not adjusting for the set ID all yielded inflated type I error. The strategy of performing unrestricted permutation led to highly inflated type I error, which is the opposite of its performance in testing a binary trait.

For testing (7) the interaction in matched-pair data, the strategy of performing unrestricted permutation yielded extremely conservative type I error. Figure 2.7 confirmed the lack of power with unrestricted permutation.

The power and sensitivity of various 1:m matched study designs were contrasted in Figures 2.8–2.10. Figure 2.8 showed that the matched-pair design always gained substantial efficiency over an analysis of data from an equivalent number of *independent* cases and controls over a wide range of θ_2 values. Figure 2.9 shows that, with a fixed number of total samples, maximizing the number of distinct sets (i.e., using 1:1 pairs) rather than increasing the number of controls per set optimized efficiency. In Figure 2.10, we show that adding more control samples to each set, while keeping the number of sets fixed, has a relatively small effect on power and sensitivity; the addition of each successive control sample yielded diminishing returns. Taken together, Figures 2.8-2.10 suggest that when data have a matched structure, a matched analysis outperforms an unmatched analysis and, in general, increasing the number of controls in a 1:m matched study beyond 1:2 may only result in fairly small improvements in power and sensitivity.

2.3.3 Analysis of the MsFLASH data

The data for our first example were extracted from the study "Menopause Strategies: Finding Lasting Answers for Symptoms and Health" (MsFLASH) (Mitchell et al., 2017; Joffe et al., 2014). This double-blinded, randomized trial enrolled women into one of three-arms: oral estradiol (arm 1), oral venlafaxine (arm 2) (two commonly used drugs to alleviate menopausal hot flashes) or placebo (arm 3). To examine the effect of these drugs on the vaginal microbiome, 113 vaginal swab samples were collected at baseline (before treatment), and at weeks 4 and 8 post-treatment. 16S rRNA gene sequencing was performed, and the results were summarized into 171 OTUs. Specifically, 9 sets (women) in the estradiol arm, 10 in the venlafaxine arm, and 18 in the placebo arm have data from swab samples at all three visits; one woman in the estradiol arm only provided samples at baseline and week 4. Due to the small sample size, we also considered an enlarged "treatment" group that combined the estradiol and venlafaxine arms. The ordination plot (Figure 2.11) showed that the samples from the same woman tended to cluster together.

In each arm, we tested whether the composition of the vaginal microbiome changed between baseline and week 4, baseline and week 8, and weeks 4 and 8; each of these tests was based on 1:1 paired data. We also tested the microbiome differences pre- and post-treatment by comparing baseline and post-treatment (both week 4 and week 8) samples without differentiating between time since treatment using a 1:2 matched-set design; the estradiol arm was an exception, as one set had only two samples, resulting in unbalanced data. We applied the LDM (using the omnibus test) and `permanovaFL` (using the Bray-Curtis distance) with the proposed strategy. As a comparison, we also applied DESeq2 (adjusting for the set ID) and the Wilcoxon signed-rank test to 1:1 matched data.

We limited our analysis for each arm to OTUs that were present at least 5 times in each of the four subsets of samples, which resulted in, for example, 31 OTUs in the venlafaxine arm. All results were summarized in Table 2.2. Only the comparisons within the venlafaxine arm yielded some significant p-values (< 0.05). In particular, the LDM generated p-value 0.033 for comparing the baseline and week 4 samples, followed by a smaller p-value 0.0042 for the baseline and week 8 samples, and then the smallest p-value 0.0003 for the baseline and the combined week 4 and week 8 samples. These p-values suggested an effect of venlafaxine on the vaginal microbiome, which was strengthened at week 8 relative to week 4. However, the differences between week 4 and week 8 were not found to be significant (the LDM p-value $= 0.76$). The results of `permanovaFL` corroborated these conclusions. In the comparison of the baseline vs. weeks 4 and 8, the LDM detected four OTUs (*Campylobacter*, *Gardnerella vaginalis*, *Porphyromonas*, and *Aerococcus christensenii*) to be differentially abundant at the nominal FDR 20% (we chose a relatively high nominal FDR because of the small number of sets), whereas DESeq2 detected none and the Wilcoxon test was not applicable.

Motivated by the likely trend of strengthened effect of venlafaxine over time, we reanalyzed the data at weeks 0, 4, and 8 in the venlafaxine arm, treating "week" as a quantitative variable. However, this analysis yielded *less* significant global p-values (0.043 by the LDM and 0.096 by `permanovaFL`), suggesting that the change in OTU frequencies as a function of time since treatment initiation is probably non-linear. We also tested whether the effect of venlafaxine is the same for the five white and four black women (excluding one women in the "other" race category), i.e., we tested the interaction between week (coded as 0 vs. 4 & 8) and race. The global p-values are 0.44 by the LDM and 0.39 by `permanovaFL`, suggesting no racial difference in the effect of vanlafaxine. These non-significant p-values may also be due to the generally low power for testing interactions.

2.3.4 Analysis of the Alzheimer's disease data

The data for our second example were generated from a pair-matched study comparing the gut microbiome of 25 patients with Alzheimer's disease (AD) and their age- and sex-matched controls (Vogt et al., 2017). A covariate of particular interest was the APOE $\epsilon4$ genotype, which was coded as carriers (one or two $\epsilon4$ alleles) vs. non-carriers (zero $\epsilon4$ alleles). APOE $\epsilon4$ genotype is a potential confounder of the association between the gut microbiome and AD, as it is distributed differently in the AD patients than in the controls (AD: 72% carriers; control: 20% carriers; p-value<0.001) in the study sample, and has been found to influence the gut microbiome (Tran et al., 2019). Since matching on APOE $\epsilon4$ genotype was not used in the study design, it should be adjusted for in the association test. The microbiome data were summarized into 972 OTUs, of which 723 were present at least 5 times in the study sample and included in our analysis. We applied the same methods as those for the MsFLASH data, except we do not report results for the Wilcoxon signed-rank test, which is not applicable in the presence of a within-pair covariate.

The results were summarized in Table 2.3. Without adjustment of the APOE $\epsilon 4$ genotype, the LDM yielded p-value 0.0001 for testing the community-level association and detected 66 OTUs (at nominal FDR 10%) that were differentially abundant between AD patients and controls. After adjustment for APOE $\epsilon 4$ genotype, the LDM yielded p-value 0.0159 and detected no OTUs. The results of `permanovaFL` corroborated this conclusion. These results suggest that much of the association seen without adjusting for APOE $\epsilon 4$ genotype is due to confounding.

2.4 Discussion

We have developed a novel strategy that extends PERMANOVA (implemented in both `adonis2` and `permanovaFL`) and the LDM for analyzing matched-set microbiome data, that can account for complex design features such as unbalanced data, sample-level confounding covariates, and continuous traits of interest. This strategy corresponds to a specific application of PERMANOVA and the LDM, without modifying any of their methodologies. Our simulations show that the proposed strategy was the most efficient among all strategies we considered, when applied to either PERMANOVA or the LDM. The LDM was also superior to existing methods, such as DESeq2 and the Wilcoxon signed-rank test, for testing individual OTUs with matched-set data. In addition, our simulation studies suggested that the 1:1 matched-pair study is the most efficient design as it maintains a good balance between sequencing cost and statistical power.

Our results in analysis of the MsFLASH data did not agree with those reported by Zhao et al. (Zhao et al., 2018), who found significant effects in the "treated" group only (rather than the venlafaxine arm). Their method was based on log-ratio-transformed frequency data and used a pseudo count value of 0.01 for zero count data, which essentially resulted in a different hypothesis being tested than that used in our methods. Similarly, we found that much of the association reported by (Vogt et al., 2017) between AD disease status and the microbiome may be due to confounding by APOE $\epsilon4$ genotype. This finding emphasizes the need to develop and use microbiome methods, such as those we have reported here, that can account for complex design features, like matching with within-set confounding covariates, that are often found in epidemiological studies involving the microbiome.

Hu and Satten (Hu and Satten, 2020) have shown that for independent case-control samples, the power of the LDM was sensitive to the OTU data scale, i.e. if untransformed frequency scale or arcsin-root-transformed data were used. We found (Figure A.1) that these patterns persisted in the analysis of matched-set data. As a result, we reiterate the recommendation in Hu and Satten (2020) and use the *omnibus* test for the LDM, which corresponds to the minimum of the p-values obtained on the frequency and arcsin-root-transformed scales.

The strategy we have proposed here is applicable to any matched-set microbiome data as long as model residuals can be assumed to have an exchangeable correlation structure. In some settings, longitudinal microbiome data that have time-varying traits (i.e., time, antibiotic intake, or infection) can be reasonably assumed to have an exchangeable correlation structure. The simple within-cluster permutation approach used here is not valid for other correlation structures such as the autoregressive model. We are currently developing methods for analysis of clustered or longitudinal microbiome data having an arbitrary residual correlation structure.

Our simulation studies showed that matched-set sampling, when available, can result in a substantial increase in power to detect global associations and sensitivity to detect individual OTUs when our approach is used. This is presumably because the overdispersion parameter for the matched data is smaller than it is for independent data sampled from the same population. In the independent data sample, the overdispersion parameter describing each observation is effectively the sum of the between- and within-set heterogeneity parameters (θ_1

and θ_2 in our simulations). In the matched data, the between-set heterogeneity (represented by θ_1 in our simulations) is effectively conditioned out. Thus, we expect the advantage of a matched analysis over an unmatched analysis to increase as the between-set heterogeneity increases. Presumably when the within-set heterogeneity is large compared to the between-set heterogeneity, a matched analysis would have a smaller advantage.

2.5 Conclusions

We proposed a new strategy, i.e., including set indicator variables as covariates and permuting within sets, that can be used with both PERMANOVA and the LDM for analyzing matched-set microbiome data. These methods not only have superior performance than existing methods but can also handle many complex design features in matched-set studies such as unequal set sizes, within-set confounding covariates, and continuous traits of interest. Given the availability of proper analytical tools, future microbiome studies should preferably adopt the matched-set design to enjoy its good power as the large microbiome heterogeneity as well as most confounding factors between sets (e.g., individuals) are conditioned out.

Table 2.1: Type I error for testing the community-level hypothesis at level 0.05

Scenario and Analysis Strategy	`permanovaFL`	`adonis2`	LDM
(1) Matched-pair data			
Proposed	0.0491	0.0450	0.0479
Not adjusting for ID	0.0491	0.0441	0.0479
Unrestricted permutation	0.0024	0.0855	0.0169
(2) Unbalanced data			
Proposed	0.0471	0.0434	0.0505
Not adjusting for ID	0.0501	0.0456	0.0500
Unrestricted permutation	0.0039	0.0732	0.0280
(3) Matched-pair data with a sample-level confounder X^{sam}			
Proposed	0.0452	0.0429	0.0476
Not adjusting for X^{sam}	0.0688	0.0631	0.0800
Not adjusting for ID	0.0713	0.0328	0.0872
Unrestricted permutation	0.0016	0.0810	0.0163
(4) Matched-pair data with a set-level covariate X^{set}			
Proposed	0.0510	0.0433	0.0481
Not adjusting for ID	0.0510	0.0451	0.0481
Not adjusting for ID, adjusting for X^{set}	0.0510	0.0450	0.0481
Unrestricted permutation	0.0021	0.0874	0.0162
(5) Unbalanced data with a set-level covariate X^{set}			
Proposed	0.0479	0.0444	0.0480
Not adjusting for ID	0.0496	0.0446	0.0483
Not adjusting for ID, adjusting for X^{set}	0.0489	0.0445	0.0489
Unrestricted permutation	0.0048	0.0736	0.0257
(6) Matched-pair data with a continuous trait			
Proposed	0.0505	0.044	0.0461
Not adjusting for ID	0.0677	0.0612	0.0881
Unrestricted permutation	0.190	0.906	0.994
(7) Matched-pair data with an interaction effect			
Proposed	0.0524	0.0295	0.0536
Unrestricted permutation	0	0.0977	0

For each of the seven scenarios, results for three or four analysis strategies are presented. First in each scenario is the "Proposed" strategy that adjusts for the set ID indicators and sample-level covariates (if present), does not adjust for set-level covariates (if present), and performs restricted permutation within sets. Each alternative strategy is described by its difference from the proposed strategy; for example, "Unrestricted permutation" maintains all the elements of the proposed strategy except for replacing the recommended within-set permutation with an unrestricted permutation.

Table 2.2: Results in analysis of the MsFLASH data

					P-values for testing the community-level hypothesis		Number of OTUs detected at FDR 20%		
Arm	T1	T2	*n*-set	*n*-sam	`permanovaFL`	LDM	Wilcoxon	DESeq2	LDM
Estradiol	0	4	10	20	0.26	0.25	0	0	0
	0	8	10	19	0.32	0.52	0	0	0
	0	4 & 8	10	29	0.26	0.32	NA	2	0
	4	8	10	19	0.23	0.27	0	0	0
Venlafaxine	0	4	10	20	0.1	0.033	0	0	0
	0	8	10	20	0.01	0.0042	0	0	0
	0	4 & 8	10	30	0.0022	0.0003	NA	0	4
	4	8	10	20	0.72	0.76	0	0	0
Placebo	0	4	18	36	0.99	0.97	0	0	0
	0	8	18	36	0.77	0.74	0	0	0
	0	4 & 8	18	54	0.96	0.86	NA	1	0
	4	8	18	36	0.6	0.59	0	0	0
Treated	0	4	20	40	0.6	0.61	0	0	0
(estradiol+	0	8	20	39	0.16	0.35	0	0	0
venlafaxine)	0	4 & 8	20	59	0.26	0.31	NA	0	0
	4	8	20	39	0.18	0.18	0	0	0

T1 and T2: the time points between which the samples were compared; 0: the baseline; 4 & 8: week 4 and week 8 after treatment. *n*-set and *n*-sam: number of sets (women) and number of samples involved in each analysis. NA: the Wilcoxon test was not applicable.

Table 2.3: Results in analysis of the Alzheimer's disease (AD) data

	P-value for testing the communitye-level hypothesis		Number of OTUs detected at FDR 10%	
	`permanovaFL`	LDM	DESeq2	LDM
Without adjustment of APOE	0.0001	0.0001	168	66
With adjustment of APOE	0.0069	0.0159	66	0

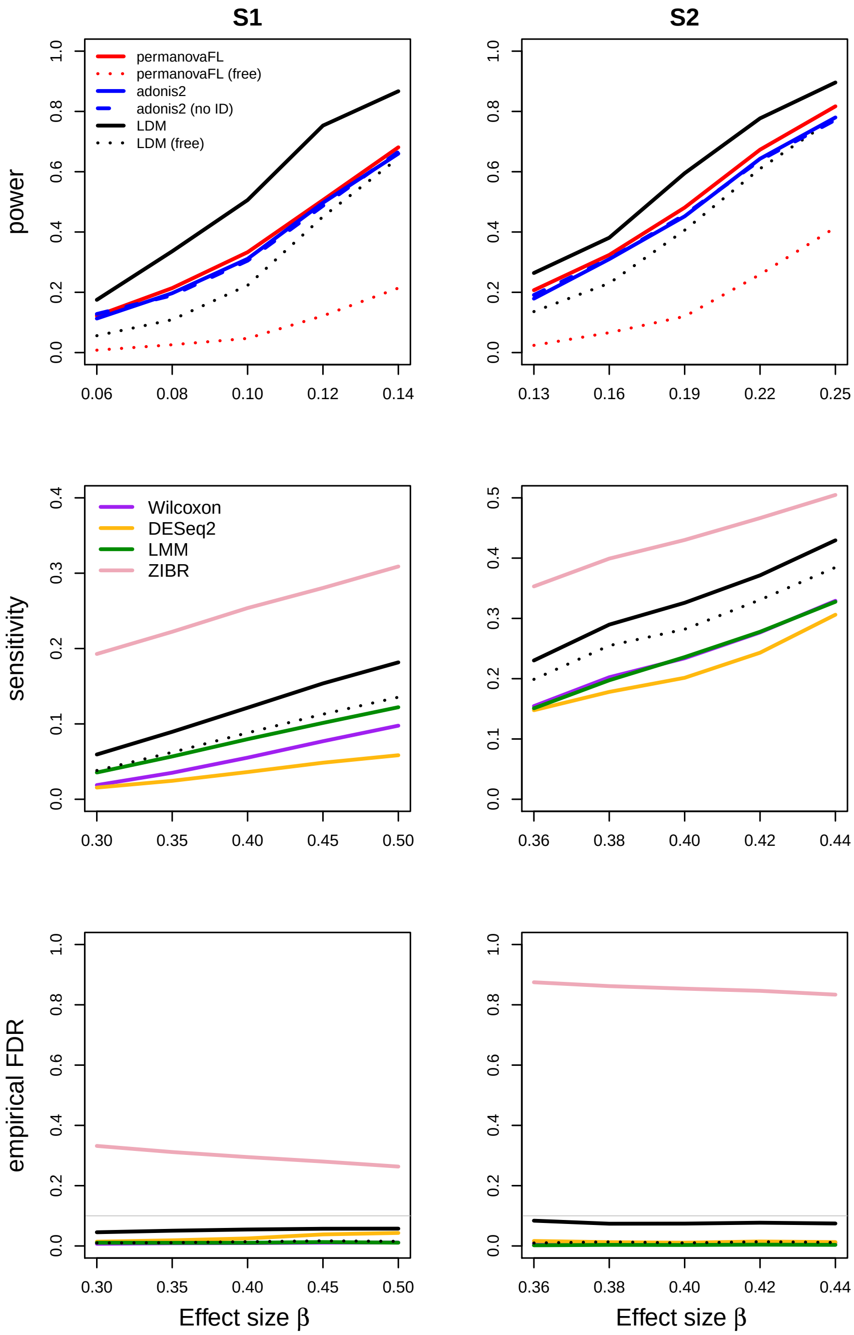


Figure 2.1: Simulation results for the matched-pair data of scenario (1). "Free" means unrestricted permutation; "no ID" means not adjusting for the set ID. Because the LDM and `permanovaFL` gave identical results with and without adjustment for set ID indicators, only results using the proposed strategy for these two methods are shown here. `adonis2` with unrestricted permutation had inflated type I error in all scenarios we examined and is therefore not shown in subsequent figures that display power or sensitivity.

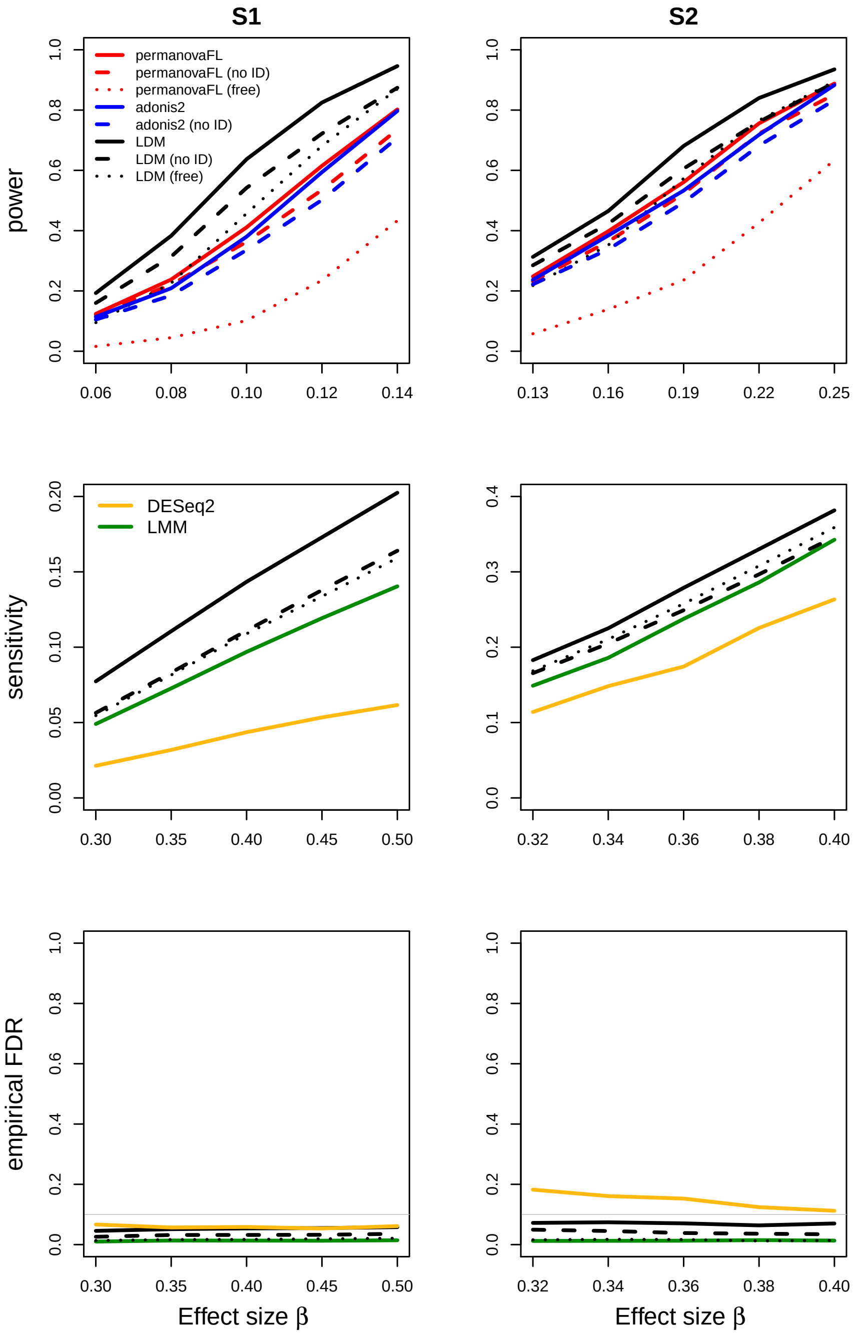


Figure 2.2: Simulation results for the unbalanced data of scenario (2).

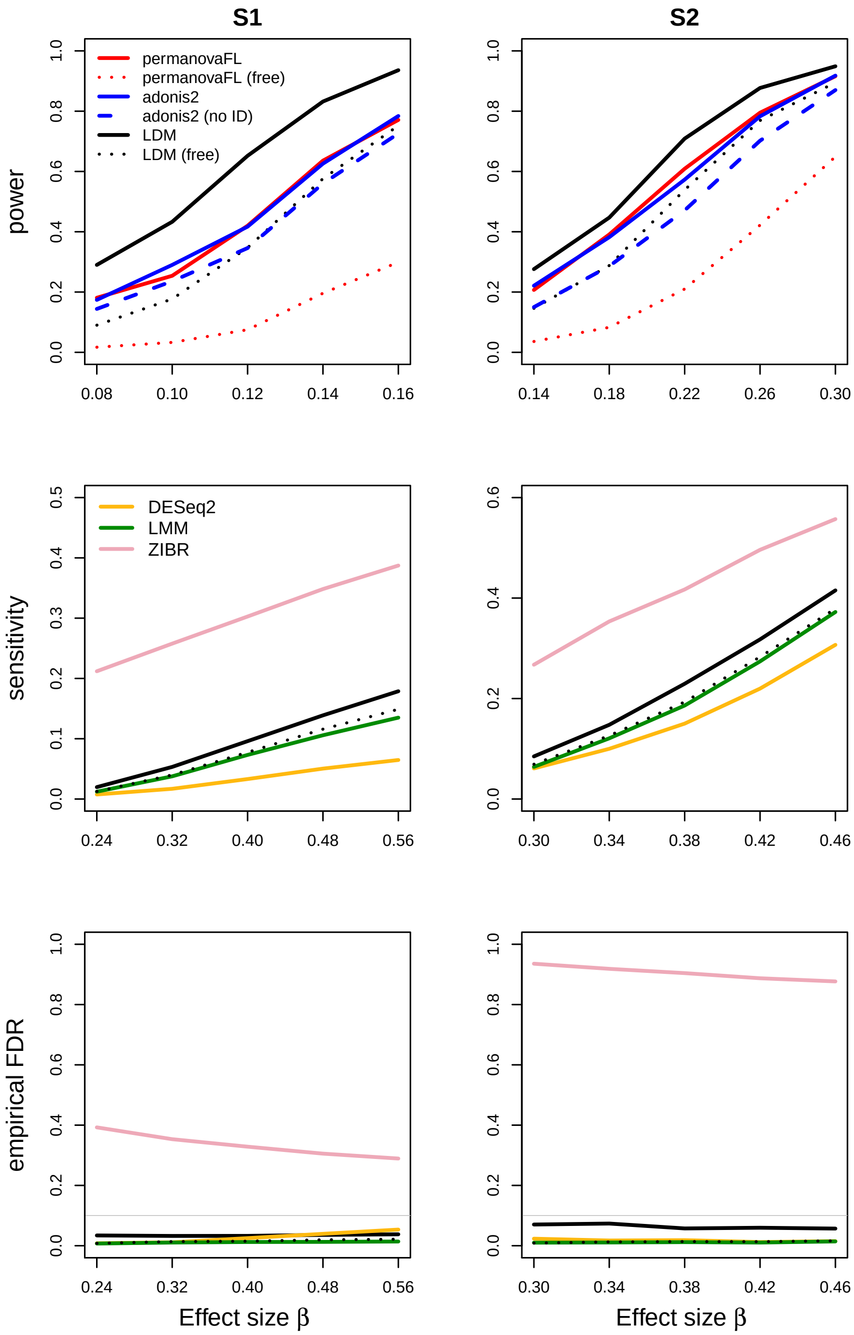


Figure 2.3: Simulation results for the matched-pair data with a sample-level confounding covariate of scenario (3). "no ID" means not adjusting for the set ID but adjusting for the confounder in this scenario. The LDM and `permanovaFL` with the "no ID" strategy had inflated type I error and thus are not shown.

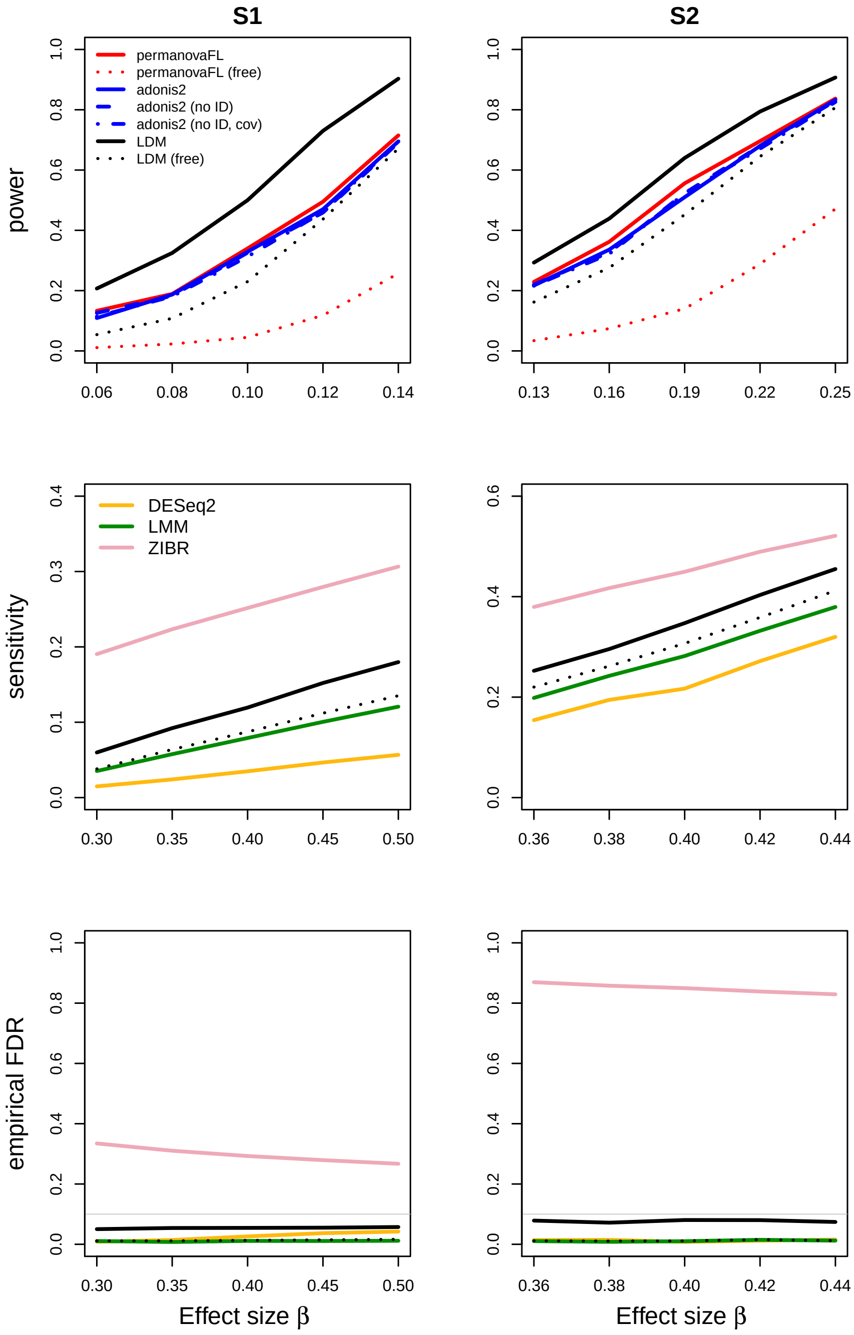


Figure 2.4: Simulation results for the matched-pair data with a set-level confounding covariate of scenario (4). "no ID" means not adjusting for the set ID (second strategy in scenario (4) of Table 2.1); "no ID, cov" means not adjusting for the set ID but adjusting for the covariate X^{set} (third strategy). The LDM and `permanovaFL` with "no ID" or "no ID, cov" had identical results as their counterparts with the proposed strategy and are thus not shown here.

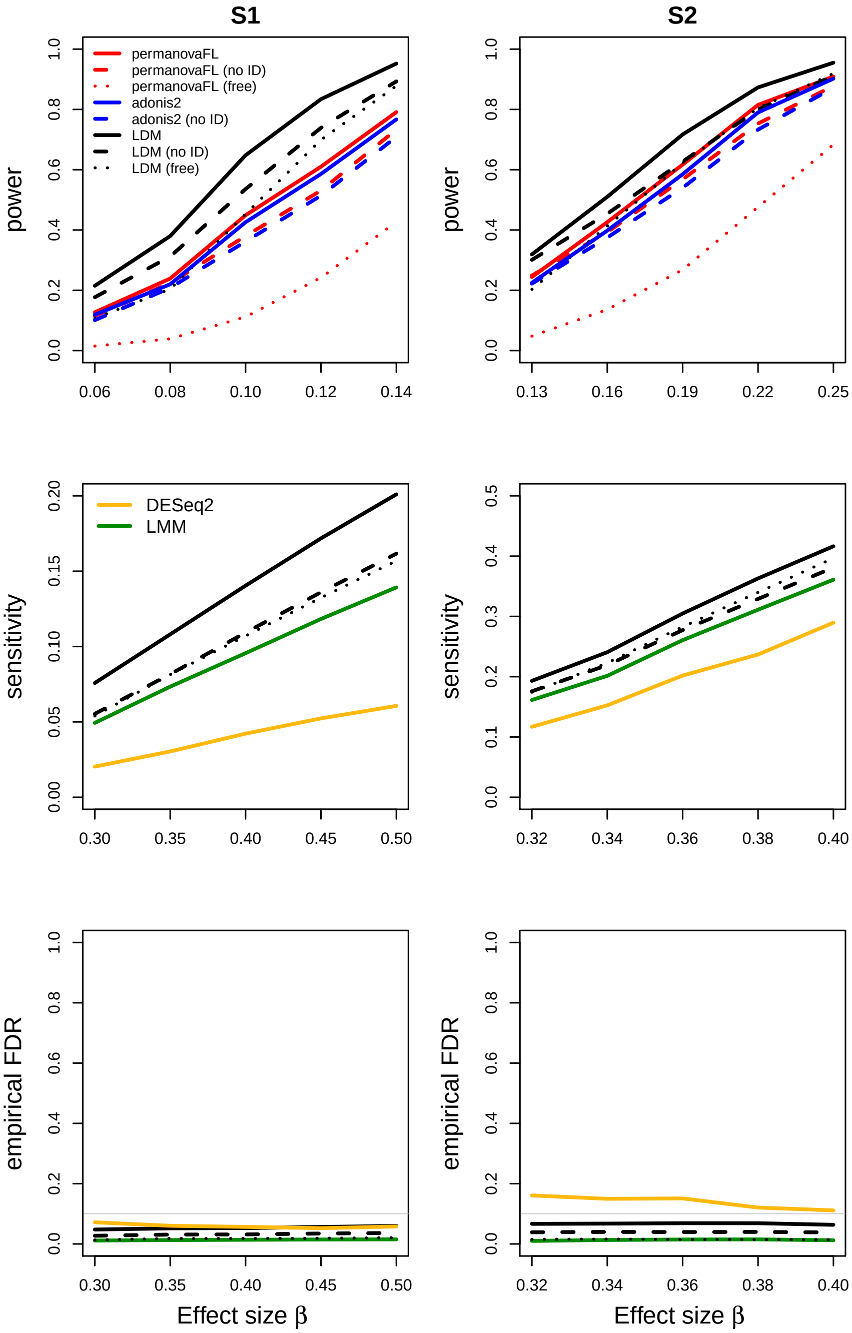


Figure 2.5: Simulation results for the unbalanced data with a set-level covariate of scenario (5). “no ID” means not adjusting for the set ID (second strategy in scenario (5) of Table 2.1). The power and sensitivity of methods using the strategy not adjusting for ID but adjusting for X^{set} (third strategy) are very similar to the power and sensitivity of their counterparts using the “no ID” strategy; thus only the latter is shown.

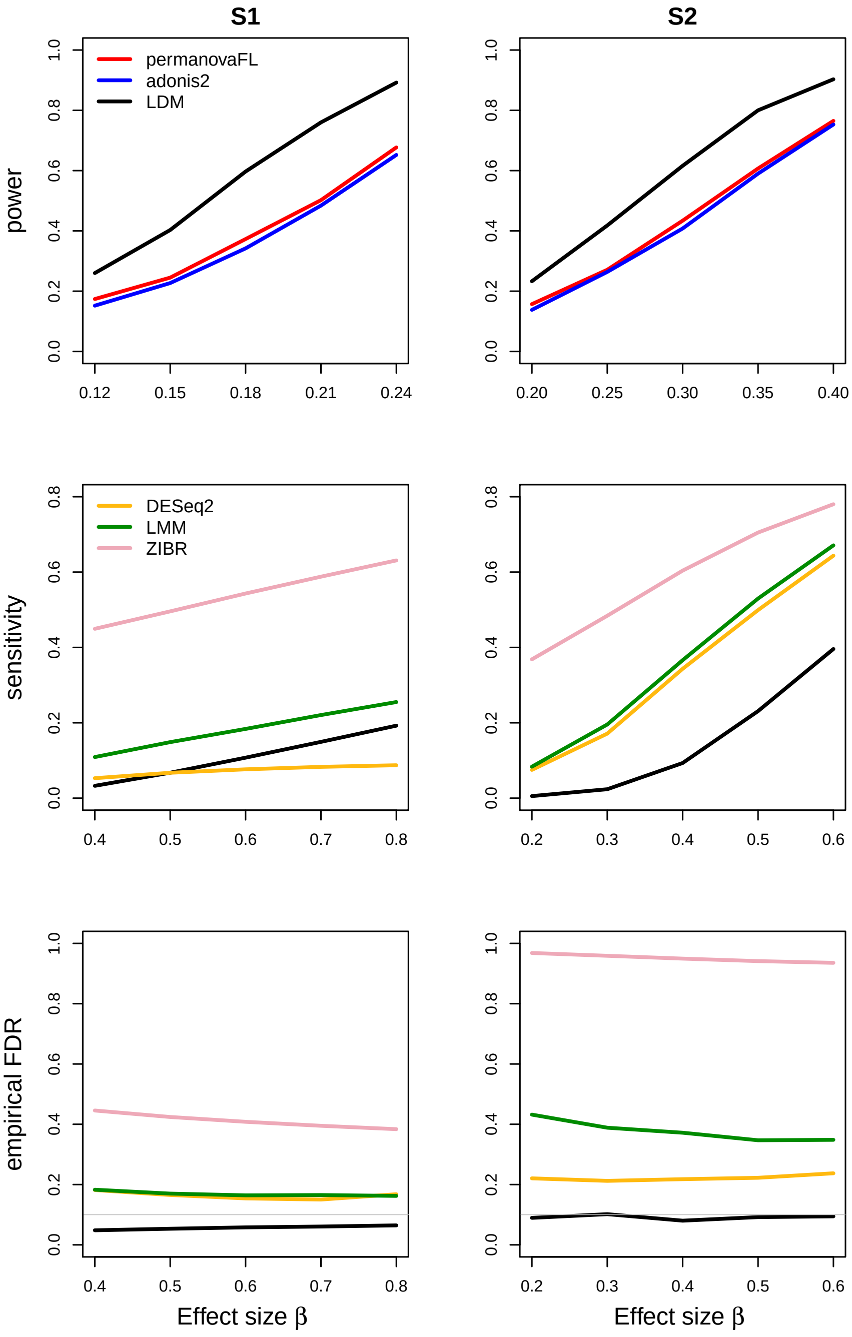


Figure 2.6: Simulation results for the matched-pair data with a continuous trait of scenario (6). The other strategies all led to inflated type I error, and are thus not shown here.

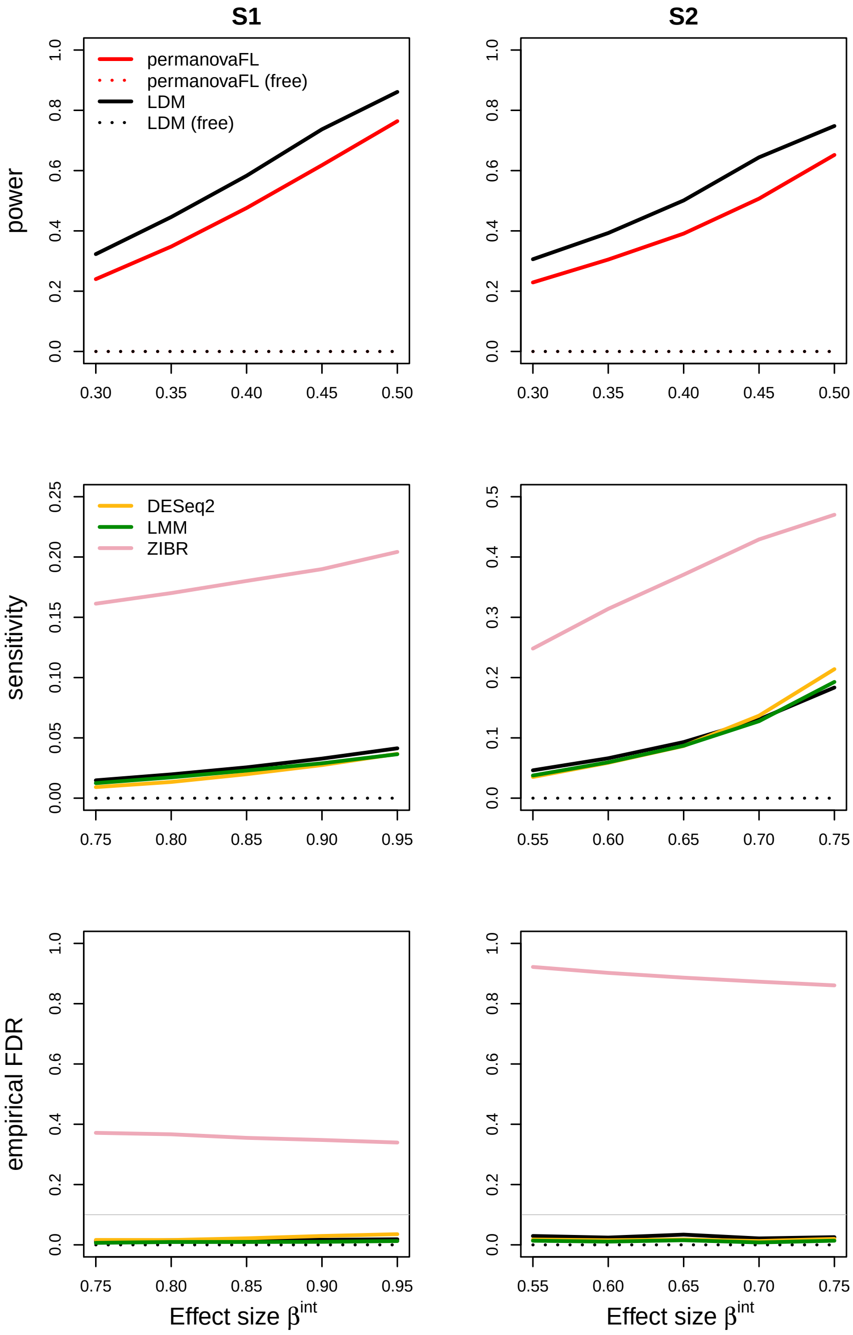


Figure 2.7: Simulation results for scenario (7) testing an interaction between a (set-level) group variable and a (sample-level) trait variable in matched-pair data.

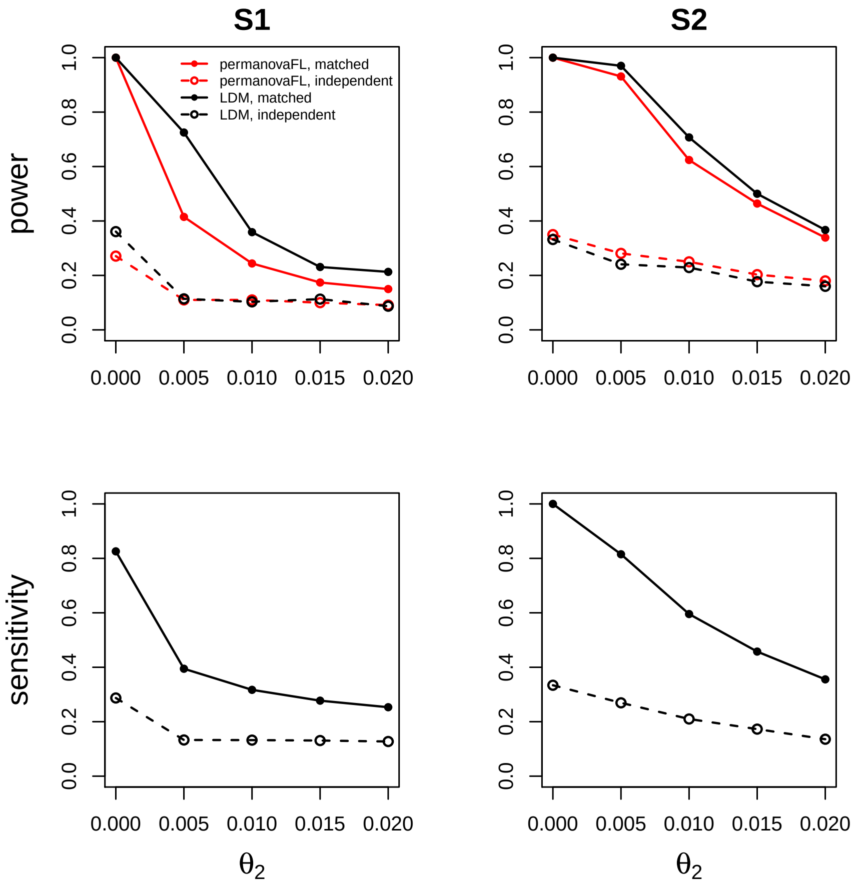


Figure 2.8: Comparing the matched-pair design (solid lines) with the independent case-control design (dashed lines) over varying sample-level heterogeneity θ_2. The effect size β was set to 0.1 (S1, power), 0.25 (S2, power), 0.8 (S1, sensitivity), and 0.6 (S2, sensitivity). All simulations shown here use between-set heterogeneity parameter $\theta_1 = 0.02$.

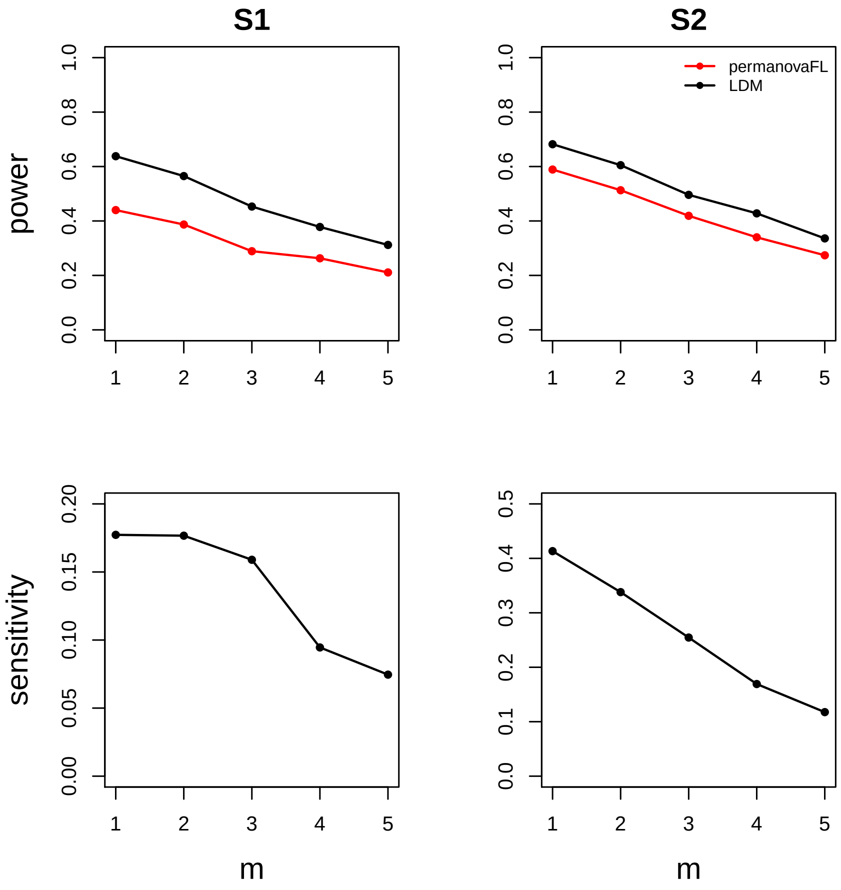


Figure 2.9: Comparing various 1:m matched-set designs with a fixed number (90) of total samples. The effect size β was set to 0.12 (S1, power), 0.22 (S2, power), 0.5 (S1, sensitivity), and 0.46 (S2, sensitivity). All simulations shown here use between-set heterogeneity parameter $\theta_1 = 0.02$ and within-set heterogeneity parameter $\theta_2 = 0.007$.

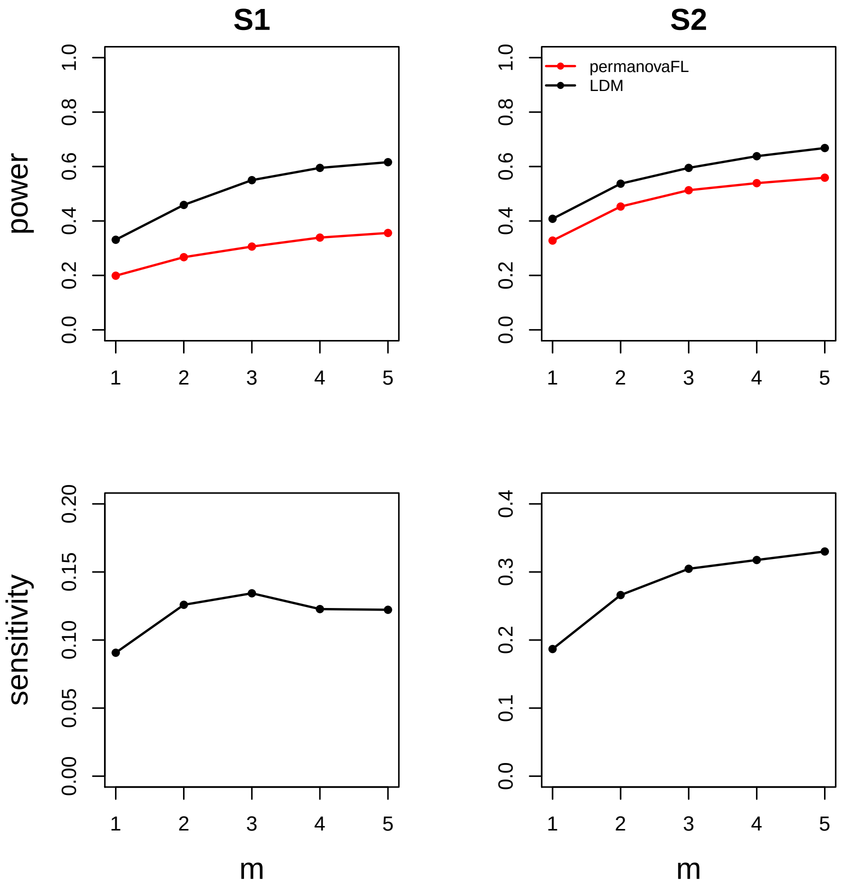


Figure 2.10: Comparing various 1:m matched-set designs with a fixed number (50) of total sets. The effect size β was set to 0.08 (S1, power), 0.16 (S2, power), 0.35 (S1, sensitivity), and 0.34 (S2, sensitivity). All simulations shown here use between-set heterogeneity parameter $\theta_1 = 0.02$ and within-set heterogeneity parameter $\theta_2 = 0.007$.

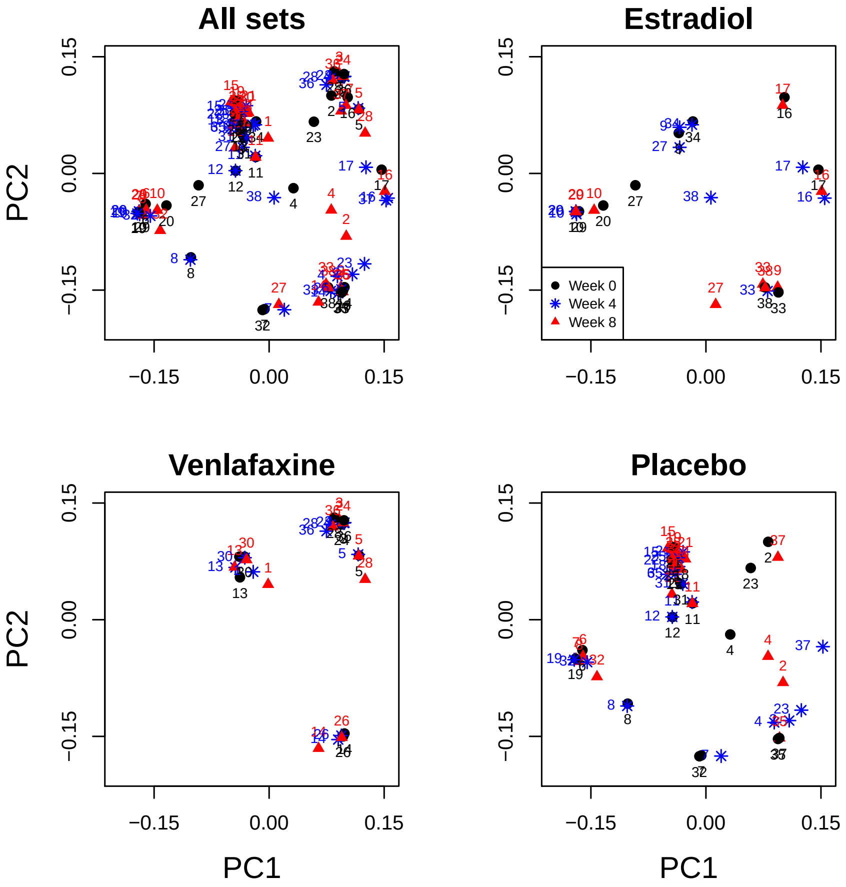


Figure 2.11: Ordination plots for the MsFLASH data. The texts above the symbols are the set IDs. The plot entitled "All sets" show the original ordination. The three plots entitled "Estradiol", "Venlafaxine", and "Placebo" show the stratified ordination by the three arms for the sake of clarity (using the same coordinates as in the plot entitled "All sets").

Chapter 3

Integrative analysis of relative abundance data and presence-absence data of the microbiome using the LDM

3.1 Introduction

LDM was initially developed for taxon data at the relative abundance scale and the arcsin-root-transformed relative abundance scale (which is variance-stabilizing for Multinomial and Dirichlet-Multinomial count data), and also offered an omnibus test that combined the results of the two taxon scales (Hu and Satten, 2020). We have shown that LDM applied to the untransformed data worked better when associated taxa were abundant and LDM applied to the transformed data worked better when associated taxa were less abundant. More recently, we made an extension of LDM for analyzing data at the presence-absence scale (Hu, Lane and Satten, 2021), which accounted for variability of library size by a rarefaction-based yet non-stochastic approach that evaluated the expected LDM test statistic over all rarefaction replicates. We found that the presence-absence analysis performed better than the relative-abundance-based analysis when associated taxa were more rare. These results motivated us to develop a *new* omnibus test for LDM that combines results from all three taxon scales. Here, we present such an omnibus test at both the community level and the individual taxon level.

3.2 Methods

3.2.1 Taxon-level omnibus test

It is straightforward to construct an LDM omnibus test for each taxon. LDM-omni, the omnibus test in Hu and Satten (2020), used the minimum of the p-values obtained from analyzing the frequency (i.e., relative abundance) data and the arcsin-root-transformed data at each taxon as the final test statistic, and used the corresponding minima from the permutation replicates to simulate the null distribution. Now we expand the test statistic to include the p-value from the presence-absence analysis in the calculation of the minimum at each taxon. As in Hu and Satten (2020), we apply Sandve's (Sandve et al., 2011) sequential

Monte-Carlo multiple-testing procedure to make discoveries with FDR control. We refer to the new omnibus test as LDM-omni3 with "3" indicates the "three" taxon scales.

Both LDM-omni and LDM-omni3 combine different scales of data at the p-value level. A completely different way to combine data would be to combine two separate lists of discoveries, each preserving FDR at some level, so that the combined list of discoveries controls the overall FDR. Kim et al. (2018) gave such a method; thus, we will compare LDM-omni3 to Kim et al.'s method that combines the discovery lists of LDM-omni and the presence-absence analysis, each using half of the overall nominal FDR. We denote this test by LDM-omni-Kim.

3.2.2 Community-level omnibus test

A community-level (global) version of LDM-omni3 could easily be constructed in the same way that LDM-omni combined information across the frequency and arcsin-root scales (Hu and Satten, 2020), by calculating an *overall* F-statistic (and corresponding p-value) for each scale (frequency, arcsin-root and, new for LDM-omni3, presence-absence) and choosing the scale with the minimum p-value. However, we also want to construct a global test that allows using the best scale *at each taxon.* Thus, we consider various p-value combination methods to combine the taxon-level LDM-omni3 p-values into a statistic we could *add* to the global LDM-omni3 test; we initially considered the minimum p-value over taxa, as well as the Cauchy (Liu and Xie, 2020), Harmonic-mean (HM) (Wilson, 2019a), Fisher's, and Stouffer's methods.

An immediate problem in combining permutation p-values calculated using B replicate datasets is that the p-values have the minimum achievable value $1/(B+1)$ (Besag and Clifford, 1991) and thus cannot well estimate the tail probability of the test statistic (Phipson and Smyth, 2010). This can greatly diminish the power of most p-value combination methods, which highly depend on the smallest p-values. To overcome this, we propose to apply the combination method to the "analytical" p-value, which is the tail probability of the F-statistic for each taxon compared to the corresponding F distribution. While the F statistics for many taxa follow the F distribution, others tend to have smaller empirical variances, in which case we scale the F statistics to have the expected variance of the F distribution. We then calculate the taxon-level "analytic" p-value for each scale of data, take the minimum of the p-values corresponding to the three scales, and combine these minimum "analytic" p-values across taxa using a p-value combination method. It is important to note here that the "analytic" p-values are not used as *real* p-values but rather a transformation of the taxon-level F-statistic to an appropriate scale for combining. The resulting statistic of a p-value combination method is assessed for significance using the permutation replicates.

For our new global LDM-omni3 test, we choose to add global tests that are based on the HM and Fisher's p-value combination methods; motivation for this choice can be found in Text S1. Thus, the global LDM-omni3 test statistic is calculated as

$$T_{\text{omni3}} = \min\{\ p_{\text{freq}},\ p_{\text{arcsin}},\ p_{\text{PA}},\ p_{\text{HM}},\ p_{\text{Fisher}}\ \},$$

where p_{freq}, p_{arcsin}, and p_{PA} are the permutation p-values of the global F-statistics at the frequency, arcsin-root and presence-absence scales as described in Hu and Satten (2020); Hu, Lane and Satten (2021), while p_{HM} and p_{Fisher} are the permutation p-values of the HM and Fisher's combinations of the "analytic" p-values. The significance of T_{omni3} is determined by permutation. We note that, by adding p-value combination tests to LDM-omni3, we increase the consistency of the taxon-level and global tests, because the taxon-level LDM-omni test in Hu and Satten (2020) reports results at the scale having the smallest p-value for that taxon, while the global LDM-omni test in Hu and Satten (2020), like p_{freq}, p_{arcsin}, and p_{PA}, is calculated using the same scale at every taxon. Finally, note that any requirement on independence of p-values when computing p-value combination tests is irrelevant here as inference is based on permutation.

3.3 Results

3.3.1 Simulation studies

Our simulations were based on the Dirichlet-Multinomial model and data on 856 taxa of the upper-respiratory-tract (URT) microbiome (Charlson et al., 2010), both of which were also used by the LDM paper (Hu and Satten, 2020). We selected five different sets of taxa to be associated with a binary or continuous trait. Ordering taxa by decreasing relative abundance, these sets are: (S1) taxa 1-10, (S2) taxa 11–50, (S3) taxa 51–200, (S4) taxa 3–5 and 11–50, and (S5) taxa 3–5 and 51–200. By design, the taxa in S1, S2, and S3 were abundant, less abundant, and relatively rare, respectively, and those in S4 and S5 were mixtures of these taxa. For each set of taxa, we considered two models for generating associations with the trait. Briefly, in Model 1, we assumed a binary trait and used different frequencies at associated taxa to simulate read count data for samples with different trait levels; in Model 2, we assumed a continuous trait and related it to associated taxa through a weighted sum of their frequencies. More detail is provided in Text S2. While Model 1 tended to create strong associations at a few taxa, Model 2 tended to simulate weak associations for all associated taxa (although the abundant taxa generally had a higher impact).

For testing individual taxa, we compared LDM-omni3 to LDM-omni and LDM-omni-Kim at the nominal FDR of 10%. To gain more insights into the relative performance of the three taxon scales, we considered the results of LDM applied to each single scale of data and call them LDM-freq, LDM-arcsin, and LDM-PA. For testing the global association, we again compared LDM-omni3 to LDM-omni. Because LDM-omni3 combined results of the five global tests, we also considered their results separately and call them LDM-freq, LDM-arcsin, LDM-PA, LDM-HM, and LDM-Fisher.

Figure 3.1 shows the results of sensitivity of taxon-specific tests and power of the global test for all methods across all ten scenarios. Note that all methods controlled the FDR or type I error (Figure B.1). For testing individual taxa, LDM-freq, LDM-arcsin, and LDM-PA

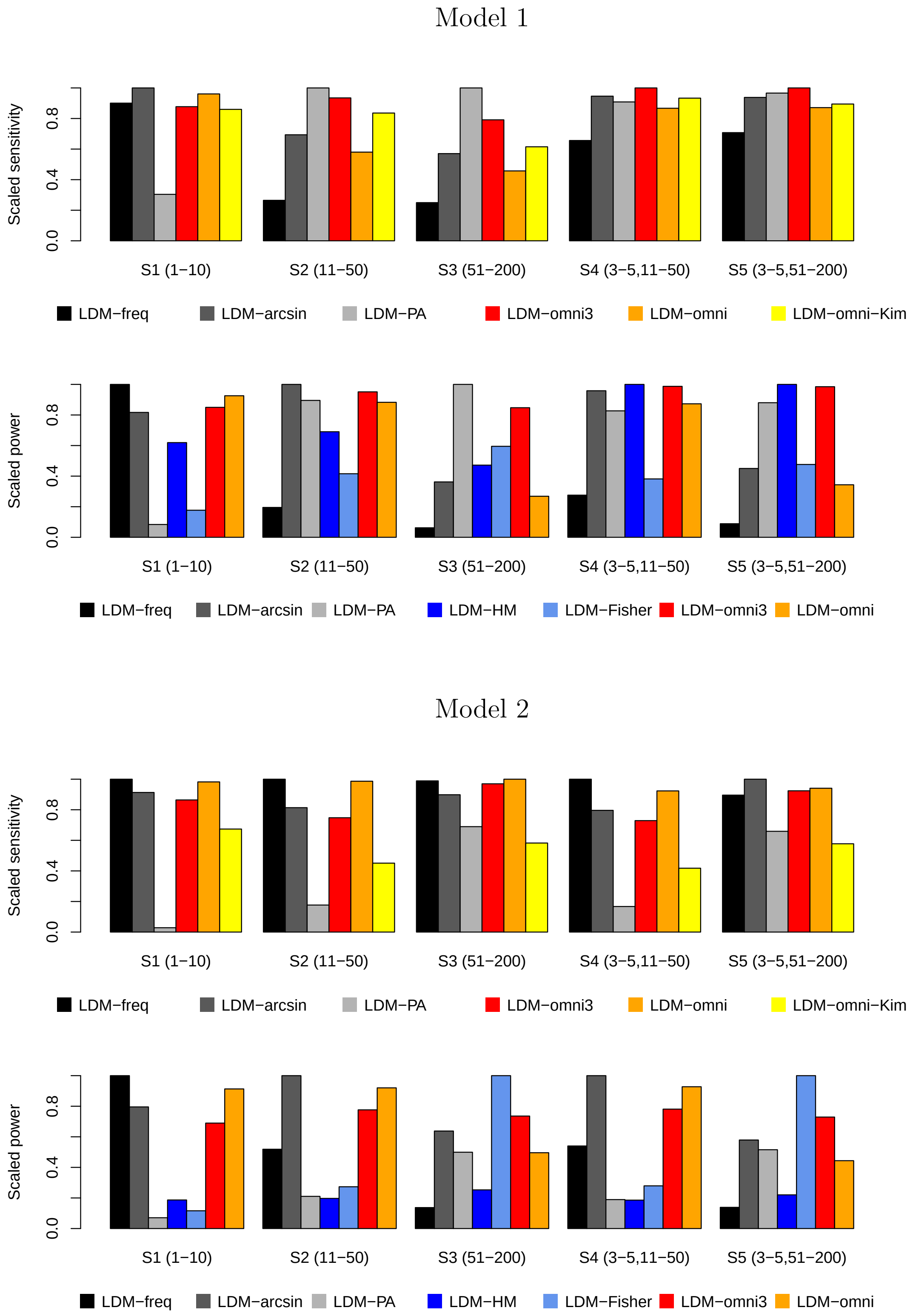


Figure 3.1: The scaled sensitivity and scaled power values were ratios against the largest value of all methods in each scenario. All results were based on 1000 replicates of data.

each performed well in some scenarios but poorly in others, whereas LDM-omni3 achieved good sensitivity across all scenarios and often tracked the best-performing scale. LDM-omni3 outperformed LDM-omni-Kim in all scenarios. LDM-omni3 yielded higher and sometimes much higher sensitivity over LDM-omni when the presence-absence analysis worked well (e.g., S2–S3 in Model 1) and lost only a small amount of sensitivity otherwise. The results at the global level showed a very similar pattern. The five tests that were components of the global LDM-omni3 test each performed well in some scenarios but poorly in others, whereas the global LDM-omni3 test achieved good power across all scenarios. When LDM-omni3 gained power over LDM-omni, the power gain can be as large as 200% (e.g., S3 and S5 in Model 1); when LDM-omni3 lost power to LDM-omni, the lost was usually small.

3.3.2 Testing Association in the URT microbiome dataset

We tested the association of the URT microbiome with smoking status in the data of Charlson et al. (Charlson et al., 2010), controlling for potential confounders gender and antibiotic use. More details on the dataset and the pre-processing procedures used can be found in Hu and Satten (2020). For testing the global association, LDM-freq, LDM-arcsin, LDM-PA, LDM-HM, and LDM-Fisher yielded p-values 0.0077, 0.00060, 0.0075, 0.020, and 0.011, respectively. LDM-omni3 yielded the second smallest p-value 0.0018, which tracked the best-performing LDM-arcsin test. In this case, when LDM-freq and LDM-arcsin worked exceptionally well, it was not surprising that LDM-omni generated the small p-value of 0.00090.

At the OTU level, LDM-freq, LDM-arcsin, and LDM-PA detected 4, 14, and 3 OTUs, respectively, that had significant associations with smoking at nominal FDR 10%. LDM-omni3 yielded 7 detections, which included two novel OTUs compared to the 5 detections by LDM-omni. One novel OTU (OTU 411) was only detected by LDM-PA. This OTU was present in 8 smokers only and absent in all others ($p = 0.0012$ by Fisher's exact test), confirming a strong association of the presence-absence data with smoking status. In contrast, the relative abundance data did not show much difference except for two smokers with fairly high values, although the difference was more pronounced on the arcsin-root scale (Figure B.2). The other novel OTU, OTU 3954, was detected by both LDM-PA and LDM-arcsin but the latter q-value was just at the boundary of the FDR threshold. This OTU was absent in 6 smokers ($p = 0.0075$ by Fisher's), also confirming a strong association at the presence-absence scale.

3.4 Conclusion

We proposed a new omnibus test within the LDM framework. In our simulations, at both the global and taxon levels, the new omnibus test tended to perform better than the initial omnibus test that was currently implemented in the LDM, especially in situations when many rare taxa were strongly associated with the covariate of interest. In the real data application, the new omnibus test made two novel detections of associated OTUs because their associations were mostly at the presence-absence scale. In summary, our new omnibus test improved over the old omnibus test. The proposed test has now been added to the LDM package, which is available on GitHub at `https://github.com/yijuanhu/LDM`. It involves little extra computational cost as the permutation replicates are already available.

Chapter 4

Estimating a Variance-Covariance Matrix with Incomplete Data

4.1 Introduction

In this article, we developed a new non-parametric method to estimate positive semi-definite variance-covariance for longitudinal data with a wide range of missingness patterns. In the methods section, we describe our algorithm and its variations in 3 more complex scenarios. In the results section, we present the simulation studies and a real microbiome longitudinal dataset that we use to assess the performance of the proposed method. We conclude with a discussion section.

4.2 Methods

Suppose that for N observations we can observe up to J potentially correlated values, and wish to calculate the $J \times J$ variance-covariance matrix. We assume that when an observation has incomplete data, we know which of the J possible values are observed. We store the data in a $n \times J$ data matrix Y. The data from the ith cluster are thus in the ith row of Y, denoted by the row vector $Y_{i\cdot}$.

It is possible to construct a version of the variance-covariance matrix element-by-element, by computing the (j, j')th element using all pairs of data having both observations j and j'. This matrix may not be positive semi-definite, so a new solution is required. Further, the nearest positive semi-definite matrix has as many zeroes as there are negative eigenvalues in the all-pairs variance-covariance estimator. Thus, we seek an estimator that utilizes the structure of the missing data problem.

4.2.1 An Estimator for Incomplete Data

Suppose that we observe K patterns of missingness. For example, if some clusters have 5 members and some have 4 members that are all missing observation 3, then there are 2 patterns of missingness. For the kth missingness pattern, let m_k denote the number of values we observe. Let D_k be the $m_k \times J$ matrix where the rth row has all zero values except for a 1 in the column corresponding to the value of j for the rth value. For the example just described, D_1 is the 5×5 identity matrix, and

$$D_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We let

$$\widehat{\Sigma}_k = \frac{1}{N_k} \sum_{i \in \mathcal{S}_k} \left(\mathcal{D}_k Y_{i\cdot}^T - \mu_k \right) \left(\mathcal{D}_k \, Y_{i\cdot}^T - \mu_k \right)^T$$

$$= \frac{1}{N_k} \sum_{i \in \mathcal{S}_k} \left(\mathcal{D}_k Y_{i\cdot}^T - \mu_k \right) \left(Y_{i\cdot} \mathcal{D}_k^T - \mu_k^T \right),$$

so that $\widehat{\Sigma}_k$ is the $m_k \times m_k$ variance-covariance matrix of observed values calculated using only observations having the kth missingness pattern. We choose μ_k to be the means of the observations in the kth missingness pattern.

Define the L_2 matrix loss function

$$L_2(V, \{\widehat{\Sigma}_k, k = 1, \cdots, K\}) = \sum_{k=1}^{K} w_k \left|\left| \widehat{\Sigma}_k V_k^{-1} - I_{m_k} \right|\right|_F^2, \tag{1}$$

where V is the $J \times J$ variance-covariance matrix for all elements, and

$$V_k = \mathcal{D}_k V \mathcal{D}_k^T,$$

is the restriction of V to the elements observed in pattern k, and where $w_k = \frac{N_k}{N}$. This choice of w_k is an attempt to achieve some efficiency.

We seek to estimate V by minimizing the L_2 matrix loss function (1). An alternative estimator would minimize the Stein loss function that is frequently used for these kinds of problems, namely

$$L_S(V, \{\widehat{\Sigma}_k, k = 1, \cdots, K\}) = \sum_{k=1}^{K} w_k \left\{ Tr\left(\widehat{\Sigma}_k V_k^{-1}\right) - ln\left[det\left(\widehat{\Sigma}_k V_k^{-1}\right)\right]\right\}, \qquad (1a)$$

where we note this loss function is intimately connected to the Normal.

4.2.2 Parameterization

Even though each $\widehat{\Sigma}_k$ is positive semi-definite, the minimizer of (1) is not guaranteed to be positive semidefinite. Thus, we must constrain the solution to consider only positive semidefinite matrices. To do this, we parameterize V using the Cholesky factorization, in which write

$$V = LL^T,$$

where L is a lower triangular matrix. For V to be positive definite, the diagonal elements of L must all be> 0. The elements of L that are in the lower triangle are not constrained. For the example considered above, with clusters having a maximum size of 5, this parameterization would be:

$$L = \begin{pmatrix} x_1 & 0 & 0 & 0 & 0 \\ x_6 & x_2 & 0 & 0 & 0 \\ x_7 & x_{10} & x_3 & 0 & 0 \\ x_8 & x_{11} & x_{13} & x_4 & 0 \\ x_9 & x_{12} & x_{14} & x_{15} & x_5 \end{pmatrix},$$

and we optimize (1) w.r.t. $\boldsymbol{x} = (x_1, \cdots, x_{15})$.

With this parameterization, it is possible to minimize (1) over positive definite matrices. We calculate the exact derivatives of (1) w.r.t. the parameters in V to take advantage of the convexity of the loss functions.

4.2.3 Convexity

We take advantage of the convexity of the objective functions we want to optimize. The convexity of L2 or Stein loss functions with respect to parameters are verified both theoretically and numerically. Firstly, the operations of matrix multiplication, determination are convex. Then, Nordström (2010) proved convexity for $Tr(C^{-1}D)$ as long as C is positive definite and D is positive semi-definite, which are the cases for $\widehat{\Sigma}_k$ and V_k^{-1} in the Stein loss function. Since the composition of two convex functions is convex, the convexity of L2/Stein loss function with respect to each parameter is established.

Leveraging loss function convexity, and manually-calculated loss function derivatives (details in Appendix), we are able to find global optimum quickly, avoiding the common speed problem due to large number of iterations before converge.

4.2.4 Different scenarios of rank and constraint

4.2.4.1 Stratum 0 (Full Data Stratum), Constraints, and Choice of Parameters

We first consider stratum 0 (the full data stratum) and consider the possible combinations of (1) the rank of $\widehat{\Sigma}_0$ and (2) the existence of known constraints when choosing our parameters.

In general, we choose to parameterize $\widetilde{V}$ (the expected variance-covariance matrix given V) since this matrix typically has full rank.

If there are no constraints, then $V = V_0 = \mathbb{V}$ ($\mathbb{V}$ is the variance-covariance matrix that describes the full data, in the absence of any constraints. Not estimable in the presence of constraints) and we could choose the parameterization $V = LL^T$. However, to handle the $k = 0$ case in the same way as the the other strata, as well as to seamlessly allow for presence of constraints, we parameterize V in the basis U (even when this is not necessary).

Case 1. We consider first the case where $\widehat{\Sigma}_0$ has full rank. Then, we can model $\widetilde{V}$ as a rank J matrix and write

$$\widetilde{V} = LL^T.$$

We then choose vec(L) to be the parameter vector. As usual, we can recover V by writing

$$V = U\widetilde{V}U^T,$$

which also gives our usual equation

$$\widetilde{V} = U^T VU.$$

These last results are all equivalent as U has full rank (and is square) in this case.

Case 2. If there are no known constraints, but $\widehat{\Sigma}_0$ has rank $m_0 < J$, then we believe the dimished rank is due to having too few observations in the full-data stratum. The notation in this case is a bit complex since the 'true' variance-covariance should have rank J but we need a rank m_0 object to compare to $\widehat{\Sigma}_0$. To be consistent with other cases, we parameterize a full rank version of $\widetilde{V}$ which we will call $\widetilde{V}_{full}$

$$\widetilde{V}_{full} = LL^T,$$

and choose vec(L) to be the parameter vector. To connect with the quantities we need for the objective function, we let U_{full} be the matrix having all the eigenvectors of $\widehat{\Sigma}_0$ (including those having eigenvalue 0), and then take

$$V_{full} = U_{full}\widetilde{V}_{full}U_{full}^T,$$

where V is the full-rank target matrix. We now want the part of V_{full} that can be represented by the columns of $U = U_0$ which we recall is the $J \times m_0$ matrix that only has the eigenvectors of $\widehat{\Sigma}_0$ that have nonzero eigenvalue. Thus, we write

$$\widetilde{V} = U^T V_{full} U = U^T U_{full} \widetilde{V}_{full} U_{full}^T U.$$

It is easy to show that if both U_{full} and U order their columns by decreasing values of the corresponding eigenvalues (this is the default) then $U^T U_{full}$ is the $m_0 \times J$ matrix given by

$$U^T U_{full} = \left(I_{m_0 \times m_0}, 0_{J-m_0 \times J-m_0} \right),$$

so that (to borrow notation from R)

$$\widetilde{V} = \widetilde{V}_{full}[1 : m_0, 1 : m_0],$$

i.e., $\widetilde{V}$ is the (1,1) block of $\widetilde{V}_{full}$ comprising the first m_0 rows and columns.

Note that in this case, it is possible that estimation of the full-rank target V will not be possible, given the observed patterns of missingness. In this case, if after estimating V we find that it is not full rank, we may want to find the eigenvectors of eigenvalue 0, and then use these to impose constraints so that we only estimate identifiable parameters. We will have to gain some experience in this situation before making a definitive recommendation.

Case 3. If there are known constraints, but $\widehat{\Sigma}_0$ attains its full rank (subject to constraints), we can only estimate

$$V = (I - C_0)\, \mathbb{V} (I - C_0)^T.$$

Thus, we will not be able to estimate the full matrix $\mathbb{V}$ in the presence of constraints. Further, V will not have full rank, and this in this situation we could not use $V = LL^T$ since V is not positive-definite.

However, if we know that $\widehat{\Sigma}_0$ has its maximum rank (i.e., if the rank of $\widehat{\Sigma}_0$ is $J-\text{rank}(C_0)$) then we can choose U_0 to be the matrix having columns given by the eigenvectors of $\widehat{\Sigma}_0$ that have non-zero eigenvalue. Then, as before, we write

$$\widetilde{V} = U^T \widehat{\Sigma} U,$$

and then write

$$\widetilde{V} = LL^T,$$

and choose $\text{vec}(L)$ to be the parameter vector.

Case 4. If there are known constrains and $\widehat{\Sigma}_0$ is itself rank-deficient (e.g. due to the full data stratum having insufficient data to estimate $\widehat{\Sigma}_0$ properly) then we must be careful when we construct U. In this situation, we must augment U so that it has $J-\text{rank}(C_0)$ columns, but we must ensure that none of the columns of U corresponds to a direction that is prohibited by the constraints.

The easiest way to do this is to let $\mathfrak{U}$ be the matrix whose columns are the eigenvectors of $\widehat{\Sigma}_0$ having eigenvalue 0; then form the matrix $\mathfrak{C} = (I - C_0)\mathfrak{U}$ which projects off the constraints from the columns of $\mathfrak{U}$; then take the SVD of $\mathfrak{C}$ and then augment the columns of U with the right singular vectors of $\mathfrak{C}$ that correspond to nonzero singular values. This will give us an extra set of columns that are linear combinations of the original eigenvectors with eigenvalue 0, but that have no component in the directions prohibited by the constraints.

4.2.4.2 All Other Strata

In the absence of constraints, previous notes already handle the case when $\widehat{\Sigma}_k$ is not full rank as everything is expressed in terms of the U_ks.

In the presence of constraints, all that is needed is to replace D_k by $\mathcal{D}_k = (I - C_k)D_k$, i.e. to apply the constraints to the expected value matrices V_k.

4.3 Results

4.3.1 Simulation and real data analysis

To generate our simulation data, we used a motivating dataset collected by Cox et al (Cox MJ, 2017), i.e., OTU data of the non-cystic fibrosis (CF) bronchiectasis microbiome in sputum. Specifically, 16S rRNA gene sequencing of the sputum microbiota was successful for 381 samples from 76 patients, resulting in 352 OTUs. The patients were followed for six months, with up to 9 measurements in the duration, accompanied with exacerbation conditions. Before analyzing the data, we removed the exacerbation measurements in between visits, removed the exacerbation measurements if the normal measurements were not missing for the same visit, resulting in 7 time points. Our goal is to estimate the variance-covariance matrix of the 7 time points on OTU relative abundance. We focused on analyzing the most abundant OTU. We also treated 0 OTU abundance as missing, removed samples with only 1 available observation, , resulting in 41 samples, 24 missingness patterns in total.

To simulate data, firstly, we applied our method on log-transformed OTU relative abundance to obtain an variance-covariance matrix estimator $\hat{\Sigma}$. This variance-covariance matrix was then used to generate multivariate normal distribution data with mean 0, and the same missingness patterns of original data was enforced on this simulated data by removing the correspondent missing values.

We compared our method to the empirical variance-covariance estimator using pair-wise available observations, under both scenarios where a constraint on the data was or was not present. To illustrate the major improvement our method provided compared to the empirical variance-covariance estimator, we focused on comparing the eigenvalues of the estimated variance-covariance matrix.

The constraints were added to simulation or original data in different ways. Specifically, for simulation data, we added the constraint that the vector (1,1,1,1,1,1,1) has eigenvalue 0, i.e. replace $\hat{\Sigma}$ by $\hat{\Sigma}_{constr} = (I - 11'/7)\hat{\Sigma}(1 - 11'/7)$ where $\hat{\Sigma}$ is the estimator we get for the full data, and then generate multivariate normal data using this $\hat{\Sigma}_{constr}$. These data will naturally have the constraint that they sum to zero. For the original real data, we centered each observation by its mean over non-missing values.

The results of estimating variance-covariance matrix for simulation and real data are summarized in Table 4.1 and Table 4.2. For simulation data, when there is no constraint, the empirical and augmented estimators yielded the same variance-covariance estimate. When the constraint is present, the empirical estimate has negative eigenvalue, thus is not positive-semidefinite; while the augmented estimate has all positive eigenvalues, thus is positive-definite, but not quite agrees with the true variance-covariance which has a 0 eigenvalue because of the constraint; the constrained augmented estimate yields satisfactory results, which agree well with the true variance-covariance matrix's eigenvalues $(6.23, 4.00, 2.69, 2.24, 0.75, 0.20, 0)$. We also calculated the Euclidean distance between the estimator and the true variance-covariance matrix to measure how close the estimate is from the truth. The Euclidean distance between the truth and the constrained augmented estimator (3.19) is smaller than that between the truth and the empirical estimator using pair-wise complete observations (4.11, after converting negative eigenvalues to 0), which confirms the superiority of the constrained augmented estimator. For the real sputum microbiome data, the empirical estimate always yields negative eigenvalues whether constraint is present or not, thus is not a qualified variance-covariance estimator. The augmented estimator always yields positive-definite variance-covariance estimate, and if constraint is present, the constrained augmented estimator has the satisfactory characteristic of being rank-deficient (positive-semidefinite).

4.4 Discussion

We have developed a novel nonparametric variance-covariance estimator that is flexible enough to handle longitudinal data with various missingness patterns and constraints.

Table 4.1: Results in analysis of the simulation and sputum microbiome data without constraint

Data	Estimator	eigenvalues of the variance-covariance estimator
Simulation	pairwise.complete.obs	(11.60, 5.51, 3.62, 2.50, 1.64, 0.55, 0.10)
	Augmented	(11.60, 5.51, 3.62, 2.50, 1.64, 0.55, 0.10)
Sputum microbiome	pairwise.complete.obs	(25.20, 6.22, 3.97, 2.65, 2.07, 0.57, -0.19)
	Augmented	(25.20, 6.22, 3.97, 2.65, 2.07, 0.57, 0.19)

Augmented refers to the proposed method using L2 loss function.

Table 4.2: Results in analysis of the simulation and sputum microbiome data with constraint

Data	Estimator	eigenvalues of the variance-covariance estimator
Simulation	pairwise.complete.obs	(5.56, 3.23, 2.92, 1.31, 0.63, 0.17, -0.046)
	Augmented	(5.56, 3.23, 2.92, 1.31, 0.63, 0.17, 0.046)
	Augmented and constrained	(5.92, 3.65, 2.91, 1.86, 0.62, 0.31 0)
Sputum microbiome	pairwise.complete.obs	(5.34, 1.59, 1.30, 0.56, 0.16, -0.026, -0.59)
	Augmented	(5.34, 1.59, 1.30, 0.59, 0.56, 0.16, 0.026)
	Augmented and constrained	(5.34, 1.59, 1.29, 0.56, 0.16, 0.12, 0)

Augmented refers to the proposed method using L2 loss function. Augmented and constrained refers to the proposed method using L2 loss function while the constraint $C_0 = 11'$ was informed to the algorithm.

We have demonstrated with simulation and real datasets that the proposed estimator is robust and always yields positive-semidefinite variance-covariance estimate, while the empirical variance-covariance estimate lacks this characteristic when multiple missingness patterns or constraint are present.

Due to the nonparametric nature of our proposed estimator, it is widely applicable in modern longitudinal studies that increasingly collect data with complex missingness patterns or constraint. The proposed method generally leads to variance-covariance estimate that is close to the truth. As such, it can greatly facilitate analysis of longitudinal studies and other research work where estimating positive-semidefinite variance-covariance matrix is necessary.

Appendix A

Appendix for Chapter 2

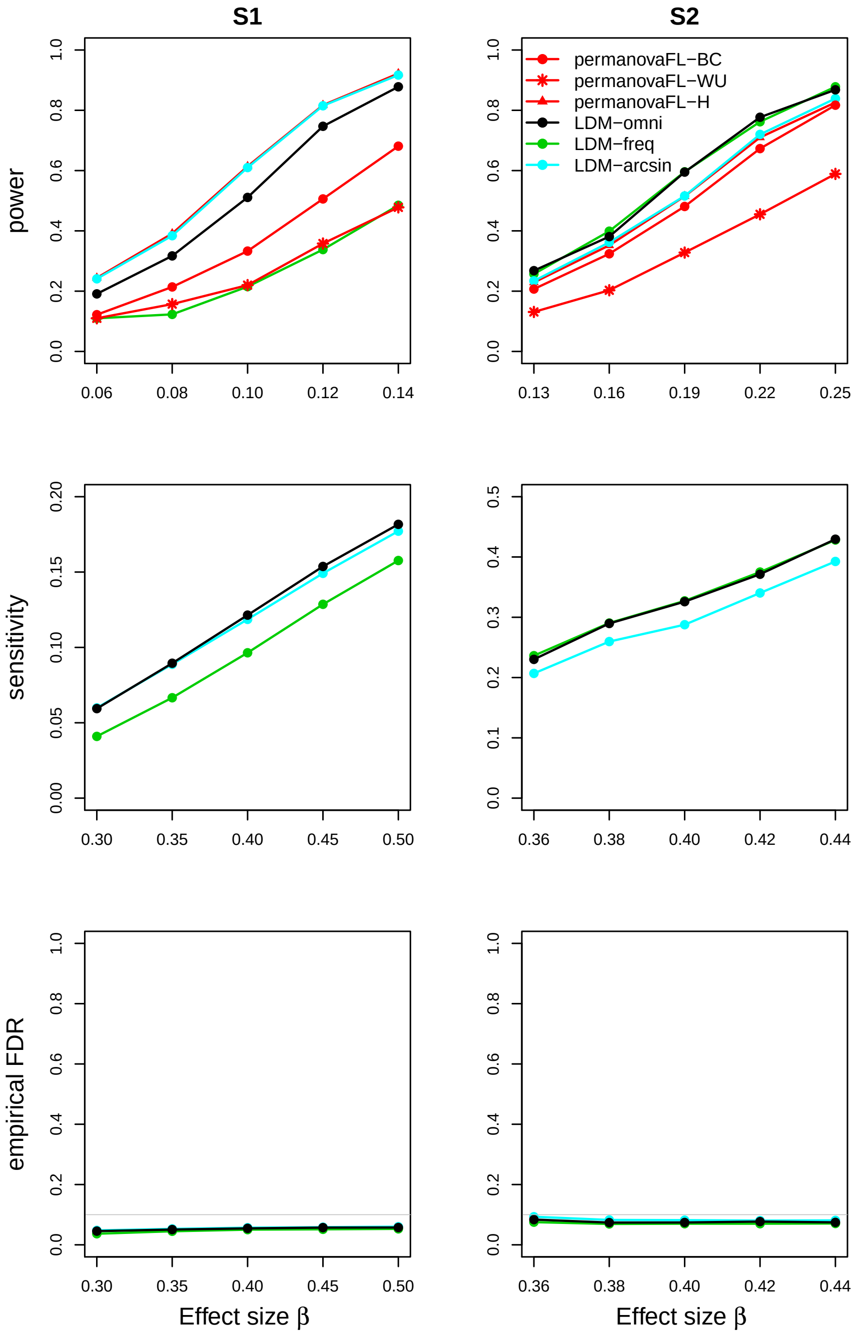


Figure A.1: Simulation results for the matched-pair data of scenario (1). The LDM on frequency scale (freq), arcsin-root transformed frequency scale (arcsin), and the omnibus test (omni) are shown. `permanovaFL` based on the Bray-Curtis (BC), weighted UniFrac (WU), and Hellinger (H) distances are shown. All methods adopted the proposed strategy.

Appendix B

Appendix for Chapter 3

B.1 Choosing among *p*-value combination methods

A new feature of LDM-omni3, compared to LDM-omni given in Hu and Satten (2020), is the use of statistics that aggregate the taxon-level omnibus p-values using a p-value combination method. The most commonly used p-value combination methods, namely, the minimum p-value, Cauchy (Liu and Xie, 2020), Harmonic-mean (HM) (Wilson, 2019a), Fisher's, and Stouffer's methods, have a decreasing emphasis on the smallest p-values and an increasing focus on the proportion of modest to weak signals (Loughin, 2004; Heard and Rubin-Delanchy, 2018) in the order given here. Thus, different methods suit different scenarios; more detail is provided in Table B.1.

Since each combination method has its own strength, it is desirable to further combine their results to form our new, omnibus global test. However, in our simulations, we found that: the Cauchy method always gave almost identical results as the HM method; the minimum p-value method performed uniformly worse than HM; and Stouffer's method performed uniformly worse than Fisher's; these results were shown in Figure B.3. Stouffer's method would work better than Fisher's method when "all nulls are equally false" (Loughin, 2004), but this is unrealistic in the microbiome setting. As a results, for our new global LDM-omni3 test, we chose to add global tests that are based on the HM and Fisher's p-value combination methods.

Table B.1: P-value combination methods

P-value combination method	Test statistic	Suitable scenario
Minimum p-value	$T_{\text{minP}} = \min_{j=1,\dots,J}\{p_j\}$	One strongest signal
Cauchy	$T_{\text{Cauchy}} = \sum_{j=1}^{J} \tan\{(0.5 - p_j)\pi\}$	A very few strong signals
Harmonic-mean (HM)	$T_{\text{HM}} = \sum_{j=1}^{J} p_j^{-1}$	A very few strong signals
Fisher's	$T_{\text{Fisher}} = -2\sum_{j=1}^{J} \log(p_j)$	A few strong to moderate signals
Stouffer's	$T_{\text{Stouffer}} = \sum_{j=1}^{J} \Phi^{-1}(1 - p_j/2)$	A large proportion of weak signals

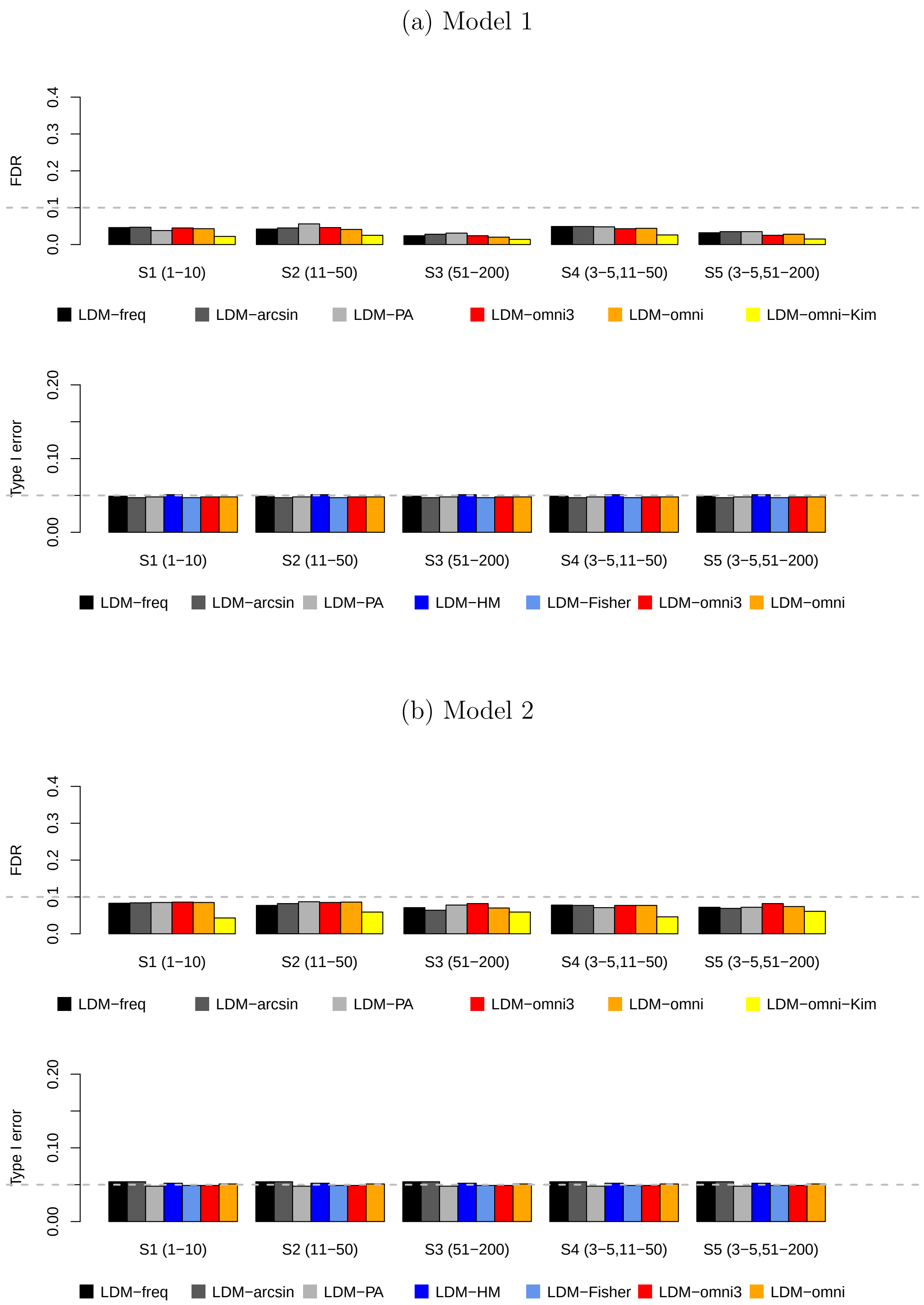
Note: $\{p_1, p_2, \dots, p_J\}$ are individual p-values. T_{HM} here is the inverse of the usual HM statistic so that T_{HM} has the property of a usual test statistic that a large value corresponds to a stronger evidence against the null hypothesis. $\Phi(\cdot)$ is the standard normal cumulative distribution function.

B.2 Two models for simulating microbiome-trait associations

Model 1 with a binary trait was previously considered in the LDM paper (Hu and Satten, 2020). We let X_i denote the trait of sample i and assumed 50 samples with $X_i = 1$ and 50 with $X_i = 0$. We let π_0 be the vector of taxon frequencies estimated from the upper-respiratory-tract microbiome data; we assign π_0 to samples for which $X_i = 0$. We derived a second set of taxon frequencies π_1 by first setting $\pi_1 = \pi_0$ and then randomly permuting the frequencies in π_1 that belonged to the selected set of taxa associated with the trait, which ensured the same frequencies in π_0 and π_1 for taxa not selected. We then defined a sample-specific frequency vector as $\widetilde{\pi}(X_i|\beta) = (1-\beta X_i)\pi_0 + \beta X_i \pi_1$, where β can be interpreted as the effect size of the trait on the overall community composition. The strengths and directions of the effects of the trait on individual taxa were heterogeneous because the resulting frequencies at each taxon were characterized not only by β but also by the differences between π_0 and π_1, which varied in magnitude and sign at different taxa. Because the frequency distribution was highly skewed towards zero, this model tended to create strong associations at only a very few taxa, leaving the majority weakly associated. Finally, we generated the taxon count data for each sample using the Dirichlet-Multinomial model with mean $\widetilde{\pi}(X_i|\beta)$, overdispersion 0.02, and library size sampled from $N(10000, (10000/3)^2)$ and left-truncated at 500.

Model 2 with a continuous trait is similar to the one considered in the MiRKAT paper (Zhao et al., 2015). We first generated the taxon count data for 100 samples using the same Dirichlet-Multinomial model as above except for the mean, which was set to π_0 here. Define $C_i = \sum_{j \in \mathcal{A}} \delta_j Y_{ij} / \overline{Y}_j$, where Y_{ij} was the observed frequency (taxon count divided by library size) of the jth taxon in the ith sample, $\overline{Y}_j$ was the average frequency for the jth taxon across samples, δ_j was randomly drawn with value 1 or -1 with equal probabilities (and fixed across replicates of data), and $\mathcal{A}$ was the set of associated taxa. Note that the direction parameter δ_j had fixed value 1 for all j in the simulations reported in Zhao et al. (2015), but was varied here to ensure (approximately) no association at taxa not selected into $\mathcal{A}$. In part, this represented our interest in testing individual taxa, which was not considered by Zhao et al. (2015). Finally, we simulated the continuous trait as $X_i = \beta \text{scale}(C_i) + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and scale(.) standardized the input vector to have mean 0 and standard deviation 1. This model tended to simulate weak associations for all associated taxa, although the abundant taxa generally had a higher impact.

Prior to analysis, we filtered out taxa that were present in fewer than 5 samples, which resulted in $\sim$460 taxa remaining in each simulated dataset. For evaluating sensitivity of detecting associated taxa, β was set to 0.5 for S1–S5 in Model 1 and 5 for S1–S5 in Model 2. These values were chosen so that the empirical sensitivity values had reached a plateau. For evaluating power of the global test, β was set to 0.2, 0.3, 0.3, 0.2, and 0.1 for S1–S5, respectively, in Model 1 and 0.5, 1, 5, 1, and 5 for S1–S5, respectively, in Model 2. These values were chosen so that the empirical power for the most powerful method in a scenario was between 70% and 90%.



(a) Model 1

(b) Model 2

Figure B.1: The gray dashed lines represent the nominal FDR 10% or the nominal type I error 0.05. The empirical FDR and type I error results were based on 1000 and 10000 replicates of data, respectively.

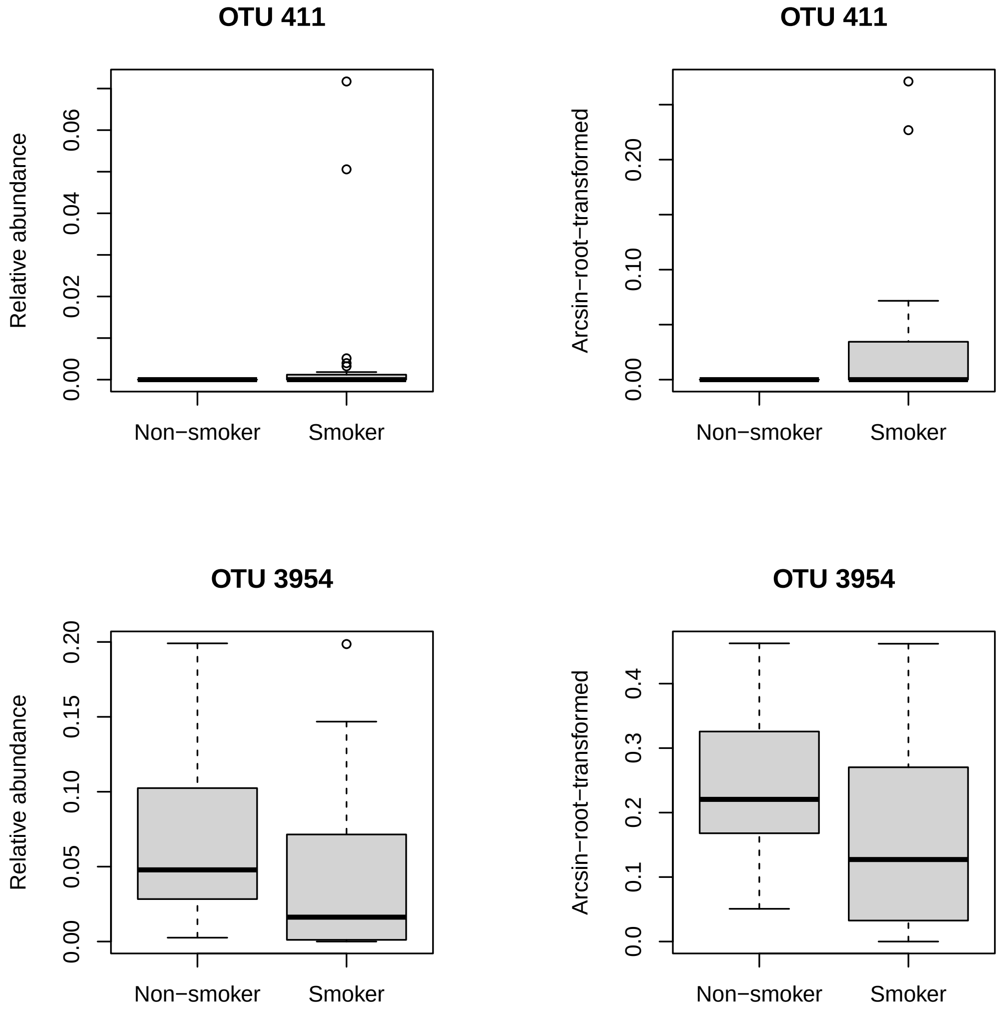
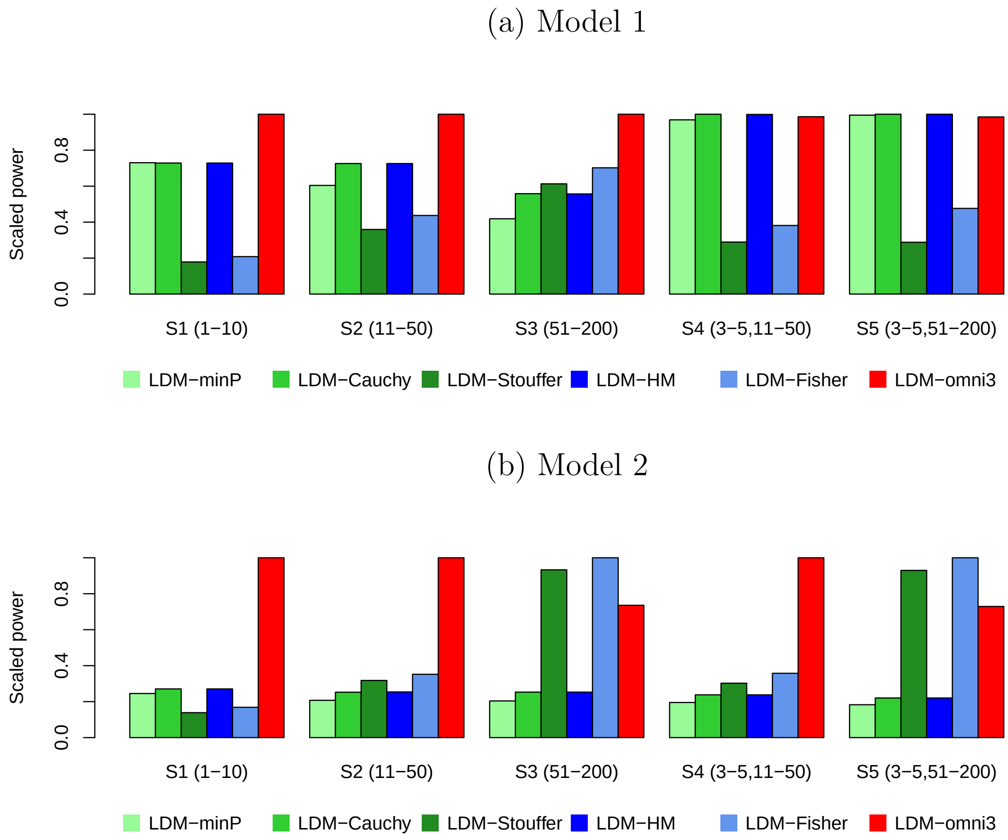


Figure B.2: Distributions of the relative abundance data and the arcsin-root-transformed data of the two taxa, OTU 411 and OTU 3954, that were detected by LDM-omni3 but not by LDM-omni in analysis of the upper-respiratory-tract microbiome data.


(a) Model 1

(b) Model 2

Figure B.3: The results of LDM-HM, LDM-Fisher, and LDM-omni3 are the same as those in Figure 2, except that they were scaled against the largest power of the methods considered here.

Appendix C

Appendix for Chapter 4

C.1 Derivatives of the Loss Function

C.1.1 Derivatives of the Stein Loss Function

We use the chain rule, first taking derivatives with respect to V, and then using the fact that $V = LDL^T$, to obtain derivatives respect to each parameters, details can be found in Appendix.

$$\frac{\partial L}{\partial x_\ell} = \sum_{i,j} \frac{\partial L}{\partial V_{ij}} \frac{\partial V_{ij}}{\partial x_\ell} = \sum_{i,j} \mathcal{U}_{ij} \frac{\partial V_{ij}}{\partial x_\ell}$$

where

$$\frac{\partial}{\partial V} \sum_k w_k \left\{ Tr\left(\widehat{\Sigma}_k V_k^{-1}\right) - ln\left(det\left[\widehat{\Sigma}_k V_k^{-1}\right]\right)\right\}$$

$$= \sum_k w_k \, D_k^T V_k^{-1} \left\{ V_k - \widehat{\Sigma}_k \right\} V_k^{-1} D_k := \mathcal{U}$$

For the first n elements of x (the diagonals of V) we then have

$$\frac{\partial L}{\partial x_\ell} = D_\ell \, (L^T \mathcal{U} L^T)_{\ell\ell} \quad , \quad 1 \le \ell \le J$$

The remaining elements can be written as

$$\frac{\partial L}{\partial x_\ell} = \sum_{i,j} \mathcal{U}_{ij} \left\{ \delta \left[i = i(\ell) \right] D_{j(\ell)} L_{jj(\ell)} + \delta \left[j = i(\ell) \right] D_{i(\ell)} L_{ii(\ell)} \right\}$$

$$= 2 \left[\mathcal{U} L D + (\mathcal{U} L D)^T \right]_{i(\ell),j(\ell)} \quad , \quad J + 1 \le \ell \le M$$

C.1.2 Derivatives of L2 Loss Function

The derivative of the L2 loss function is nearly identical to that of the Stein loss.

$$\frac{\partial \, ||\widehat{\Sigma}_k V_k^{-1} - I||^2}{\partial V} = \frac{\partial \, Tr \, \left(\widehat{\Sigma}_k V_k^{-1} - I \right) \left(V_k^{-1} \widehat{\Sigma}_k - I \right)}{\partial V}$$

$$= -D_k^T V_k^{-1} \widehat{\Sigma}_k \left\{ \widehat{\Sigma}_k V_k^{-1} - I \right\} V_k^{-1} D_k - D_k^T V_k^{-1} \left\{ V_k^{-1} \widehat{\Sigma}_k - I \right\} \widehat{\Sigma}_k V_k^{-1} D_k$$

which is symmetric (note the 2nd term on the RHS is the transpose of the 1st term on the RHS).

Bibliography

Aitchison, J. (1982), 'The statistical analysis of compositional data', Journal of the Royal Statistical Society: Series B (Methodological) **44**(2), 139–160.

Aitchison, J. (1986), The statistical analysis of compositional data, Chapman and Hall, London-New York.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000), 'Logratio analysis and compositional distance', Mathematical Geology **32**(3), 271–275.

Aitchison, J. and Ho, C. (1989), 'The multivariate poisson-log normal distribution', Biometrika **76**(4), 643–653.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', Journal of Molecular Biology **215**(3), 403–410.

Anderson, M. J. (2001), 'A new method for non-parametric multivariate analysis of variance', Austral ecology **26**(1), 32–46.

Anderson, M. J. and Legendre, P. (1999), 'An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model', Journal of statistical computation and simulation **62**(3), 271–303.

Asher, J. E., Lamb, J. A., Brocklebank, D., Cazier, J.-B., Maestrini, E., Addis, L., Sen, M., Baron-Cohen, S. and Monaco, A. P. (2009), 'A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12', The American Journal of Human Genetics **84**(2), 279–285.

Bai, J., Hu, Y. and Bruner, D. (2019), 'Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7–18 years old children from the american gut project', Pediatric Obesity **14**(4), e12480.

BECG, C. B. and Gray, R. (1984), 'Calculation of polychotomous logistic regression parameters using individualized regressions', Biometrika **71**(1), 11–18.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', Journal of the royal statistical society. Series B (Methodological) pp. 289–300.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H. et al. (2020), 'Microbiome definition re-visited: old concepts and new challenges', Microbiome **8**(1), 1–22.

Berkson, J. (1944), 'Application of the logistic function to bio-assay', Journal of the American Statistical Association **39**(227), 357–365.

Berkson, J. (1953), 'A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function', Journal of the American Statistical Association **48**(263), 565–599.

Besag, J. and Clifford, P. (1991), 'Sequential Monte Carlo p-values', Biometrika **78**(2), 301–304.

Bezdek, J. C. and Hathaway, R. J. (2003), 'Convergence of alternating optimization', Neural, Parallel & Scientific Computations **11**(4), 351–368.

Boca, S., Heller, R. and Sampson, J. (2018), 'Multimed: Testing multiple biological mediators simultaneously', R package version **2**(0).

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C. and Sampson, J. N. (2014), 'Testing multiple biological mediators simultaneously', Bioinformatics **30**(2), 214–220.

Bogomolov, M. and Heller, R. (2018), 'Assessing replicability of findings across two studies of multiple features', Biometrika **105**(3), 505–516.

Bokulich, N. A., Dillon, M. R., Zhang, Y., Rideout, J. R., Bolyen, E., Li, H., Albert, P. S. and Caporaso, J. G. (2018), 'q2-longitudinal: longitudinal and paired-sample analyses of microbiome data', MSystems **3**(6).

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E., Da Silva, R., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G., Lee, J., Ley, R., Liu, Y., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson II, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J. G. (2018), QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science, Technical report, PeerJ Preprints.

Breslow, N. E., Day, N. E., Davis, W. et al. (1980), Statistical methods in cancer research: volume 1-the analysis of case-control studies, Vol. 32, IARC.

Brill, B., Amir, A. and Heller, R. (2019), 'Testing for differential abundance in compositional counts data, with application to microbiome studies', arXiv **XX**(XX).

Brooks, J. P. (2016), 'Challenges for case-control studies with microbiome data', Annals of epidemiology **26**(5), 336–341.

Burns, M. B., Lynch, J., Starr, T. K., Knights, D. and Blekhman, R. (2015), 'Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment', Genome Medicine **7**(1), 55.

Cai, T. T. and Zhang, A. (2016), 'Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data', Journal of Multivariate Analysis **150**, 55–74.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016), 'Dada2: high-resolution sample inference from illumina amplicon data', Nature methods **13**(7), 581.

Chao, A. and Chiu, C.-H. (2016), 'Bridging the variance and diversity decomposition approaches to beta diversity vis similarity and differentiation measures', Methods in Ecology and Evolution **7**(8), 919–928.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D. and Collman, R. G. (2010), 'Disordered microbial communities in the upper respiratory tract of cigarette smokers', PloS one **5**(12), e15216.

Chen, E. Z. and Li, H. (2016), 'A two-part mixed-effects model for analyzing longitudinal microbiome compositional data', Bioinformatics **32**(17), 2611–2617. PMCID: PMC5860434.

Chen, J. and Chen, L. (2017), 'Gmpr: A novel normalization method for microbiome sequencing data', bioRxiv p. 112565.

Chen, J. and Li, H. (2013), 'Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis', The annals of applied statistics **7**(1).

Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., Chandra, S., McGarrell, D. M., Schmidt, T. M., Garrity, G. M. and Tiedje, J. M. (2003), 'The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy', Nucleic Acids Research **31**(1), 442–443.

Cox, M. J., Cookson, W. O. and Moffatt, M. F. (2013), 'Sequencing the human microbiome in health and disease', Human molecular genetics **22**(R1), R88–R94.

Cox MJ, Turek EM, H. C. M. G. J. P. C. M. e. a. (2017), 'Longitudinal assessment of sputum microbiome by sequencing of the 16s rrna gene in non-cystic fibrosis bronchiectasis patients', PLoS ONE **12**(2), e0170622.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006), 'Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb', Applied and Environmental Microbiology **72**(7), 5069–5072.

Dolan, K. T. and Chang, E. B. (2017), 'Diet, gut microbes, and the pathogenesis of inflammatory bowel diseases', Molecular nutrition & food research **61**(1), 1600129.

Dunlop, A. L., Satten, G. A., Hu, Y.-J., Knight, A. K., Hill, C. C., Wright, M. L., Smith, A. K., Read, T. D., Pearce, B. D. and Corwin, E. J. (2021), 'Vaginal microbiome composition in early pregnancy and risk of spontaneous preterm and early term birth among african american women', Frontiers in Cellular and Infection Microbiology **11**.

Everson, R. (1998), 'Orthogonal, but not orthonormal, procrustes problems', Advances in computational Mathematics **3**(4).

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R. and Gloor, G. B. (2014), 'Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis', Microbiome **2**(1), 15. PMCID: PMC4030730.

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', Biometrika **80**(1), 27–38.

Freedman, D. and Lane, D. (1983), 'A nonstochastic interpretation of reported significance levels', Journal of Business & Economic Statistics **1**(4), 292–298.

Friston, K. J., Penny, W. D. and Glaser, D. E. (2005), 'Conjunction revisited', Neuroimage **25**(3), 661–667.

Gandy, A. and Hahn, G. (2014), 'MMCTest-a safe algorithm for implementing multiple Monte Carlo tests', Scandinavian Journal of Statistics **41**(4), 1083–1101.

Gandy, A. and Hahn, G. (2016), 'A framework for Monte Carlo based multiple testing', Scandinavian Journal of Statistics **43**(4), 1046–1063.

Gandy, A. and Hahn, G. (2017), 'QuickMMCTest: quick multiple Monte Carlo testing', Statistics and Computing **27**(3), 823–832.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. and Egozcue, J. J. (2017), 'Microbiome datasets are compositional: and this is not optional', Frontiers in microbiology **8**, 2224.

Gower, J. C. (1966), 'Some distance properties of latent root and vector methods used in multivariate analysis', Biometrika **53**(3-4), 325–338.

Guo, W. and Peddada, S. (2008), 'Adaptive choice of the number of bootstrap samples in large scale multiple testing', Statistical applications in genetics and molecular biology **7**(1).

Guo, X., Pan, W., Connett, J. E., Hannan, P. J. and French, S. A. (2005), 'Small-sample performance of the robust score test and its modifications in generalized estimating equations', Statistics in medicine **24**(22), 3479–3495.

Haaland, R. E., Fountain, J., Hu, Y., Holder, A., Dinh, C., Hall, L., Pescatore, N. A., Heeke, S., Hart, C. E., Xu, J. et al. (2018), 'Repeated rectal application of a hyperosmolar lubricant is associated with microbiota shifts but does not affect pr ep drug concentrations: results from a randomized trial in men who have sex with men', Journal of the International AIDS Society **21**(10), e25199.

Haldane, J. (1956), 'The estimation and significance of the logarithm of a ratio of frequencies', Annals of human genetics **20**(4), 309–311.

Hamidi, B., Wallace, K. and Alekseyenko, A. V. (2019), 'MODIMA, a method for multivariate omnibus distance mediation analysis, allows for integration of multivariate exposure-mediator-response relationships', Genes **10**(7), 524.

Hawinkel, S., Mattiello, F., Bijnens, L. and Thas, O. (2017), 'A broken promise: microbiome differential abundance methods do not control the false discovery rate', Briefings in bioinformatics **20**(1), 210–221.

Heard, N. A. and Rubin-Delanchy, P. (2018), 'Choosing between methods of combining p-values', Biometrika **105**(1), 239–246.

Higham, N. J. (2002), 'Computing the nearest correlation matrix-a problem from finance', IMA Journal of Numerical Analysis **22**, 329–343.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', Scandinavian journal of statistics pp. 65–70.

Hu, J., Koh, H., He, L., Liu, M., Blaser, M. J. and Li, H. (2018), 'A two-stage microbial association mapping framework with advanced FDR control', Microbiome **6**(1), 131. PMCID: PMC6060480.

Hu, Y. J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E. and Lin, D.-Y. (2013), 'Meta-analysis of gene-level associations for rare variants based on single-variant statistics', *American Journal of Human Genetics* **93**(2), 236–248. PMCID: PMC3738834.

Hu, Y.-J., Lane, A. and Satten, G. A. (2021), 'A rarefaction-based extension of the ldm for testing presence-absence associations in the microbiome', *Bioinformatics* p. https://doi.org/10.1093/bioinformatics/btab012.

Hu, Y. J., Li, Y., Auer, P. L. and Lin, D. Y. (2015), 'Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations', *Proceedings of the National Academy of Sciences* **112**(4), 1019–1024. PMCID: PMC4313847.

Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A. S. and Satten, G. A. (2016), 'Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls', *PLoS genetics* **12**(5), e1006040. PMCID: PMC4859496.

Hu, Y. J., Lin, D. Y., Sun, W. and Zeng, D. (2014), 'A likelihood-based framework for association analysis of allele-specific copy numbers', *Journal of the American Statistical Association* **109**(508), 1533–1545. PMCID: PMC4315366.

Hu, Y. J., Lin, D. Y. and Zeng, D. (2010), 'A general framework for studying genetic effects and gene–environment interactions with missing data', *Biostatistics* **11**(4), 583–598. PMCID: PMC3294269.

Hu, Y.-J. and Satten, G. A. (2020), 'Testing hypotheses about the microbiome using the linear decomposition model (LDM)', *Bioinformatics* pp. bbtaa260, https://doi.org/10.1093/bioinformatics/btaa260.

Hu, Y.-J. and Satten, G. A. (2021), 'A rarefaction-without-resampling extension of permanova for testing presence-absence associations in the microbiome', *bioRxiv* p. https://doi.org/10.1101/2021.04.06.438671.

Hu, Y.-J., Sun, W., Tzeng, J.-Y. and Perou, C. M. (2015), ‘Proper use of allele-specific expression improves statistical power for cis-eqtl mapping with rna-seq data’, Journal of the American Statistical Association **110**(511), 962–974.

Hu, Y. and Lin, D. (2010), ‘Analysis of untyped snps: maximum likelihood and imputation methods’, Genetic epidemiology **34**(8), 803–815.

Hu, Y., Satten, G. A. and Hu, Y.-J. (2021), ‘Locom: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control’, bioRxiv p. https://doi.org/10.1101/2021.10.03.462964.

Hu, Y., Satten, G. A. and Hu, Y.-J. (2022), ‘Testing associations of the microbiome with censored survival outcomes using the LDM and PERMANOVA’, bioRxiv .

Hugerth, L. W. and Andersson, A. F. (2017), ‘Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing’, Frontiers in Microbiology **8**, 1561.

Hughes, J. B. and Hellmann, J. J. (2005), ‘The application of rarefaction techniques to molecular inventories of microbial diversity’, Methods in enzymology **397**, 292–308.

Jiang, H. and Salzman, J. (2012), ‘Statistical properties of an early stopping rule for resampling-based multiple testing’, Biometrika **99**(4), 973–980.

Joffe, H., Guthrie, K. A., LaCroix, A. Z., Reed, S. D., Ensrud, K. E., Manson, J. E., Newton, K. M., Freeman, E. W., Anderson, G. L., Larson, J. C. et al. (2014), ‘Low-dose estradiol and the serotonin-norepinephrine reuptake inhibitor venlafaxine for vasomotor symptoms: a randomized clinical trial’, JAMA internal medicine **174**(7), 1058–1066.

Kaul, A., Mandal, S., Davidov, O. and Peddada, S. D. (2017), ‘Analysis of microbiome data in the presence of excess zeros’, Frontiers in microbiology **8**, 2114. PMCID: PMC5682008.

Kim, Y., Lim, J., Lee, J. S. and Jeong, J. (2018), ‘Controlling two-dimensional false discovery rates by combining two univariate multiple testing results with an application to mass spectral data’, Chemometrics and Intelligent Laboratory Systems **182**, 149–157.

Kleinbaum, D. G., Kupper, L. L., Nizam, A. and Muller, K. G. (2007), Applied Regression Analysis and Other Multivariable Methods, Duxbury Press.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L., Smith, J. D., DiDonato, J. A., Chen, J., Li, H., Wu, G. D., Lewis, J. D., Warrier, M., Brown, J. M., Krauss, R. M., Tang, W. H., Bushman, F. D., Lusis, A. J. and Hazen, S. L. (2013), ‘Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis’, Nature medicine **19**(5), 576. PMCID: PMC3650111.

Koh, H., Blaser, M. J. and Li, H. (2017), ‘A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping’, Microbiome **5**(1), 45.

Koh, H., Li, Y., Zhan, X., Chen, J. and Zhao, N. (2019), ‘A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies’, Frontiers in genetics **10**, 458. PMCID: PMC6532659.

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S. and Bravo, H. C. (2018), ‘Analysis and correction of compositional bias in sparse sequencing count data’, BMC genomics **19**(1), 799.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G. and Shannon, W. D. (2012), ‘Hypothesis testing and power calculations for taxonomic-based human microbiome data’, PloS one **7**(12), e52078.

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., Stevens, H. J., Bennett, W. E., Shaikh, N., Linneman, L. A., Hoffmann, J. A., Hamvas, A., Deych, E., Shands, B. A., Shannon, W. D. and Tarr, P. (2014), ‘Patterned progression of bacterial populations in the premature infant gut’, Proceedings of the National Academy of Sciences **111**(34), 12522–12527. PMCID: PMC4151715.

Legendre, P. and Anderson, M. J. (1999), ‘Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments’, Ecological monographs **69**(1), 1–24.

Legendre, P. and Gallagher, E. D. (2001), ‘Ecologically meaningful transformations for ordination of species data’, Oecologia **129**(2), 271–280.

Legendre, P. and Legendre, L. F. (2012), Numerical ecology, Vol. 24, Elsevier.

Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012), ‘Normalization, testing, and false discovery rate estimation for rna-sequencing data’, Biostatistics **13**(3), 523–538.

Li, Y., Hu, Y.-J. and Satten, G. A. (2019), ‘A bottom-up approach to testing hypotheses that have a branching tree dependence structure, with false discovery rate control’, arXiv:1903.06850 .

Liao, P., Satten, G. A. and Hu, Y.-J. (2017a), ‘PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies’, Genetic epidemiology **41**(5), 375–387. PMCID: PMC5564424.

Liao, P., Satten, G. A. and Hu, Y.-J. (2017b), ‘Robust inference of population structure from next-generation sequencing data with systematic differences in sequencing’, Bioinformatics **34**(7), 1157–1163. PMCID: PMC6031038.

Lin, D. Y., Hu, Y. J. and Huang, B. E. (2008), ‘Simple and efficient analysis of disease association with missing genotype data’, American Journal of Human Genetics **82**(2), 444–452. PMCID: PMC2427170.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E. and Lin, X. (2019), 'Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies', The American Journal of Human Genetics **104**(3), 410–421.

Liu, Y. and Xie, J. (2020), 'Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures', Journal of the American Statistical Association **115**(529), 393–402.

Loughin, T. M. (2004), 'A systematic comparison of methods for combining p-values from independent tests', Computational statistics & data analysis **47**(3), 467–485.

Lounici, K. (2014), 'High-dimensional covariance matrix estimation with missing observations', Bernoulli **20**(3), 1029–1058.

Love, M. I., Huber, W. and Anders, S. (2014), 'Moderated estimation of fold change and dispersion for rna-seq data with deseq2', Genome biology **15**(12), 550.

Lozupone, C. A., Hamady, M., Kelley, S. T. and Knight, R. (2007), 'Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities', Applied and environmental microbiology **73**(5), 1576–1585. PMCID: PMC1828774.

Lozupone, C. and Knight, R. (2005), 'UniFrac: a new phylogenetic method for comparing microbial communities', Applied and environmental microbiology **71**(12), 8228–8235. PMCID: PMC1317376.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. and Knight, R. (2011), 'Unifrac: an effective distance metric for microbial community comparison', The ISME journal **5**(2), 169.

Luo, X., Schwartz, J., Baccarelli, A. and Liu, Z. (2021), 'Testing cell-type-specific mediation effects in genome-wide epigenetic studies', Briefings in Bioinformatics **22**(3), bbaa131.

Majumdar, A., Witte, J. S. and Ghosh, S. (2015), 'Semiparametric allelic tests for mapping multiple phenotypes: Binomial regression and Mahalanobis distance', Genetic Epidemiology **39**(8), 635–650.

Mallick, H., Tickle, T., McIver, L., Rahnavard, G., Nguyen, L., Weingart, G., Ma, S., Ren, B., Schwager, E., Subramanian, A., Paulson, J., Franzosa, E., Corrada, B. H. and Huttenhower, C. (2019), 'Multivariable association in population-scale meta'omic surveys', In Submission .

Mancl, L. A. and DeRouen, T. A. (2001), 'A covariance estimator for gee with improved small-sample properties', Biometrics **57**(1), 126–134.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. and Peddada, S. D. (2015), 'Analysis of composition of microbiomes: a novel method for studying microbial composition', Microbial ecology in health and disease **26**(1), 27663. PMCID: PMC4450248.

McArdle, B. H. and Anderson, M. J. (2001), 'Fitting multivariate models to community data: a comment on distance-based redundancy analysis', Ecology **82**(1), 290–297.

McLaren, M. R., Willis, A. D. and Callahan, B. J. (2019), 'Consistent and correctable bias in metagenomic sequencing experiments', Elife **8**.

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

Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., LaCroix, A. Z., Fiedler, T., Munch, M., Liu, C. et al. (2017), 'Vaginal microbiota and genitourinary menopausal symptoms: a cross-sectional analysis', Menopause **24**(10), 1160–1166.

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., Stempak, J. M., Gevers, D., Xavier, R. J., Silverberg, M. S. et al. (2015), 'Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease', Genome biology **16**(1), 67.

Muller, K. E. and Fetterman, B. A. (2012), Regression and ANOVA: An Integrated Approach using SAS Software, SAS Institute.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G. and Knight, R. (2013), Advancing our understanding of the human microbiome using QIIME, in 'Methods in enzymology', Vol. 531, Elsevier, pp. 371–444.

Nichols, T., Brett, M., Andersson, J., Wager, T. and Poline, J.-B. (2005), 'Valid conjunction inference with the minimum statistic', Neuroimage **25**(3), 653–660.

Nordström, K. (2010), 'Convexity of the inverse and moore–penrose inverse', Linear Algebra and its Applications **434**(2011), 1489–1512.

Nyante, S. J., Gammon, M. D., Kaufman, J. S., Bensen, J. T., Lin, D. Y., Barnholtz-Sloan, J. S., Hu, Y., He, Q., Luo, J. and Millikan, R. C. (2011), 'Common genetic variation in adiponectin, leptin, and leptin receptor and association with breast cancer subtypes', Breast Cancer Research and Treatment **129**(2), 593–606.

Olesen, S. W., Vora, S., Techtmann, S. M., Fortney, J. L., Bastidas-Oyanedel, J. R., Rodríguez, J., Hazen, T. C. and Alm, E. J. (2016), 'A novel analysis method for paired-sample microbial ecology experiments', PloS one **11**(5), e0154804.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R. and Coin, L. J. (2012), 'MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS', PloS One **7**(5).

Palarea-Albaladejo, J. and Martin-Fernandez, J. A. (2015), 'zCompositions–R package for multivariate imputation of left-censored data under a compositional approach', Chemometrics and Intelligent Laboratory Systems **143**, 85–96.

Park, S. and Lim, J. (2019), 'Non-asymptotic rate for high-dimensional covariance estimation with non-independent missing observations', Statistics Probability Letters **153**, 113–123.

Paulson, J. N., Stine, O. C., Bravo, H. C. and Pop, M. (2013), 'Differential abundance analysis for microbial marker-gene surveys', Nature methods **10**(12), 1200–1202.

Phipson, B. and Smyth, G. K. (2010), 'Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn', Statistical applications in genetics and molecular biology **9**(1).

Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018), 'The madness of microbiome: attempting to find consensus ?best practice? for 16s microbiome studies', Appl. Environ. Microbiol. **84**(7), e02627–17.

Pope, J. L., Tomkovich, S., Yang, Y. and Jobin, C. (2017), 'Microbiota as a mediator of cancer progression and therapy', Translational Research **179**, 139–154.

Potter, D. M. (2005), 'A permutation test for inference in logistic regression with small-and moderate-sized data sets', Statistics in medicine **24**(5), 693–708.

Relman, D. A. (2012), 'The human microbiome: ecosystem resilience and health', Nutrition reviews **70**(suppl_1), S2–S9.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010), 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', Bioinformatics **26**(1), 139–140. PMCID: PMC2796818.

Rune Halvorsen, Ø. (2003), 'Partitioning the variation in a plot-by-species data matrix that is related to n sets of explanatory variables', Journal of Vegetation Science **14**(5), 693–700.

Sampson, J. N., Boca, S. M., Moore, S. C. and Heller, R. (2018), 'Fwer and fdr control when testing multiple mediators', Bioinformatics **34**(14), 2418–2424.

Sandve, G. K., Ferkingstad, E. and Nygård, S. (2011), 'Sequential monte carlo multiple testing', Bioinformatics **27**(23), 3235–3241.

Satten, G. A., Kong, M. and Datta, S. (2018), 'Multisample adjusted u-statistics that account for confounding covariates', Statistics in Medicine **37**(2), 3357–3372.

Satten, G. A., Tyx, R. E., Rivera, A. J. and Stanfill, S. (2017), 'Restoring the duality between principal components of a distance matrix and linear combinations of predictors, with application to studies of the microbiome', PloS one **12**(1), e0168131.

Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data, CRC Press.

Schulfer, A. F., Schluter, J., Zhang, Y., Brown, Q., Pathmasiri, W., McRitchie, S., Sumner, S., Li, H., Xavier, J. B. and Blaser, M. J. (2019), 'The impact of early-life sub-therapeutic antibiotic treatment (stat) on excessive weight is robust despite transfer of intestinal microbes', The ISME journal **13**(5), 1280–1292.

Shade, A., Peter, H., Allison, S. D., Baho, D., Berga, M., Bürgmann, H., Huber, D. H., Langenheder, S., Lennon, J. T., Martiny, J. B., Matulich, K. L., Schmidt, T. M. and Handelsman, J. (2012), 'Fundamentals of microbial community resistance and resilience', Frontiers in microbiology **3**, 417. PMCID: PMC3525951.

Shi, P. and Li, H. (2017), 'A model for paired-multinomial data and its application to analysis of data on a taxonomic tree', Biometrics **73**(4), 1266–1278.

Sohn, M. B. and Li, H. (2019), 'Compositional mediation analysis for microbiome studies', The Annals of Applied Statistics **13**(1), 661–681.

Sohn, M. B., Lu, J. and Li, H. (2021), 'A compositional mediation model for a binary outcome: Application to microbiome studies', Bioinformatics .

Sonnenburg, J. L. and Bäckhed, F. (2016), 'Diet–microbiota interactions as moderators of human metabolism', Nature **535**(7610), 56. PMCID: PMC5991619.

Stewart, C. J., Ajami, N. J., O?Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., Metcalf, G. A. et al. (2018), 'Temporal development of the gut microbiome in early childhood from the teddy study', Nature **562**(7728), 583.

Sullivan, P. F., de Geus, E. J., Willemsen, G., James, M. R., Smit, J. H., Zandbelt, T., Arolt, V., Baune, B. T., Blackwood, D., Cichon, S., Coventry, W., Domschke, K., Farmer, A., Fava, M., Gordon, S., He, Q., Heath, A., Heutink, P., Holsboer, F., Hoogendijk, W., Hottenga, J., Hu, Y., Kohli, M., Lin, D., Lucae, S., Macintyre, D., Maier, W., McGhee, K., McGuffin, P., Montgomery, G., Muir, W., Nolen, W., Nthen, M., Perlis, R., Pirlo, K., Posthuma, D., Rietschel, M., Rizzu, P., Schosser, A., Smit, A., Smoller, J., Tzeng, J., van, Dyck, R., Verhage, M., Zitman, F., Martin, N., Wray, N., Boomsma, D. and Penninx, B. (2008), 'Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo', Molecular Psychiatry **14**(4), 359–375.

Tang, Z.-Z., Chen, G. and Alekseyenko, A. V. (2016), 'Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances', Bioinformatics **32**(17), 2618–2625.

Tang, Z.-Z., Chen, G., Alekseyenko, A. V. and Li, H. (2016), 'A general framework for association analysis of microbial communities on a taxonomic tree', Bioinformatics **33**(9), 1278–1285.

Tang, Z.-Z., Chen, G., Hong, Q., Huang, S., Smith, H. M., Shah, R. D., Scholz, M. B. and Ferguson, J. F. (2019), 'Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites', Frontiers in genetics **10**, 454.

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., Sørensen, S., Bisgaard, H. and Waage, J. (2016), 'Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies', Microbiome **4**(1), 62. PMCID: PMC5123278.

Tran, T. T., Corsini, S., Kellingray, L., Hegarty, C., Le Gall, G., Narbad, A., Müller, M., Tejera, N., O?toole, P. W., Minihane, A.-M. et al. (2019), 'Apoe genotype influences the gut microbiome structure and function in humans and mice: relevance for alzheimer's disease pathophysiology', The FASEB Journal pp. fj–201900071R.

Tyx, R. E., Stanfill, S. B., Keong, L. M., Rivera, A. J., Satten, G. A. and Watson, C. H. (2016), 'Characterization of bacterial communities in selected smokeless tobacco products using 16s rdna analysis', PLoS One **11**(1), e0146939.

van Winkel, R., Rutten, B. P., Peerbooms, O., Peuskens, J., van Os, J. and De Hert, M. (2010), 'MTHFR and risk of metabolic syndrome in patients with schizophrenia', Schizophrenia Research **121**(1), 193–198.

VanderWeele, T. J. and Shpitser, I. (2011), 'A new criterion for confounder selection', Biometrics **67**(4), 1406–1413.

VanderWeele, T. J. and Vansteelandt, S. (2009), 'Conceptual issues concerning mediation, interventions and composition', Statistics and its Interface **2**(4), 457–468.

VanderWeele, T. and Vansteelandt, S. (2014), 'Mediation analysis with multiple mediators', Epidemiologic methods **2**(1), 95–115. PMCID: PMC4287269.

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J.-X. et al. (2018), 'The human gut microbiome in early-onset type 1 diabetes from the teddy study', Nature **562**(7728), 589–594.

Verbeek, E. C., Bakker, I. M., Bevova, M. R., Bochdanovits, Z., Rizzu, P., Sondervan, D., Willemsen, G., De Geus, E. J., Smit, J. H., Penninx, B. W., Boomsma, D., Hoogendijk, W. and Heutink, P. (2012), 'A fine-mapping study of 7 top scoring genes from a GWAS for major depressive disorder', PloS One **7**(5), e37384.

Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K. et al. (2017), 'Gut microbiome alterations in alzheimer's disease', Scientific reports **7**(1), 13537.

Wang, C., Hu, J., Blaser, M. J. and Li, H. (2019), 'Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data', Bioinformatics p. doi: 10.1093/bioinformatics/btz565.

Wang, K. (2014), 'Testing genetic association by regressing genotype over multiple phenotypes', PloS One **9**(9).

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R. and Knight, R. (2017), 'Normalization and microbial differential abundance strategies depend upon data characteristics', Microbiome **5**(1), 27. PMCID: PMC5335496.

Werft, W. and Benner, A. (2010), 'glmperm: A permutation of regressor residuals test for inference in generalized linear models', The R Journal **2**(1), 39–43.

Westfall, P. H. and Young, S. S. (1993), Resampling-based multiple testing: Examples and methods for p-value adjustment, Vol. 279, John Wiley & Sons.

Wijesinha, A., Begg, C. B., Funkenstein, H. H. and McNeil, B. J. (1983), 'Methodology for the differential diagnosis of a complex data set: a case study using data from routine ct scan examinations', Medical Decision Making **3**(2), 133–154.

Wilson, D. J. (2019a), 'The harmonic mean p-value for combining dependent tests', Proceedings of the National Academy of Sciences **116**(4), 1195–1200.

Wilson, D. J. (2019b), 'The harmonic mean p-value for combining dependent tests', Proceedings of the National Academy of Sciences **116**(4), 1195–1200.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014), 'Permutation inference for the general linear model', Neuroimage **92**, 381–397.

Witkin, S. and Linhares, I. (2017), 'Why do lactobacilli dominate the human vaginal microbiota?', BJOG: An International Journal of Obstetrics & Gynaecology **124**(4), 606–611.

Wu, B. and Pankow, J. S. (2015), 'Statistical methods for association tests of multiple continuous traits in genome-wide association studies', Annals of Human Genetics **79**(4), 282–293.

Wu, C., Chen, J., Kim, J. and Pan, W. (2016), 'An adaptive association test for microbiome data', Genome medicine **8**(1), 56.

Yinglin Xia, Jun Sun, D.-G. C. (2018), Statistical Analysis of Microbiome Data with R, Springer.

Yue, Y. and Hu, Y.-J. (2021), 'A new approach to testing mediation of the microbiome using the ldm', bioRxiv .

Zhai, J., Knox, K., Twigg III, H. L., Zhou, H. and Zhou, J. J. (2019), 'Exact variance component tests for longitudinal microbiome studies', Genetic epidemiology **43**(3), 250–262. PMCID: PMC6416054.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C. and Chen, J. (2017), 'A small-sample multivariate kernel machine test for microbiome association studies', Genetic epidemiology **41**(3), 210–220.

Zhang, H., Chen, J., Li, Z. and Liu, L. (2019), 'Testing for mediation effect with application to human microbiome data', Statistics in Biosciences pp. 1–16.

Zhang, J., Wei, Z. and Chen, J. (2018), ‘A distance-based approach for testing the mediation effect of the human microbiome’, Bioinformatics **34**(11), 1875–1883.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A., Zhuang, W. and Yi, N. (2018), ‘Negative binomial mixed models for analyzing longitudinal microbiome data’, Frontiers in microbiology **9**, 1683. PMCID: PMC6070621.

Zhang, Y., Han, S. W., Cox, L. M. and Li, H. (2017), ‘A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study’, Genetic epidemiology **41**(8), 769–778.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M. C. (2015), ‘Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test’, The American Journal of Human Genetics **96**(5), 797–807.

Zhao, N., Zhan, X., Guthrie, K. A., Mitchell, C. M. and Larson, J. (2018), ‘Generalized hotelling’s test for paired compositional data with application to human microbiome studies’, Genetic epidemiology **42**(5), 459–469.

Zhu, Z., Satten, G. A., Mitchell, C. and Hu, Y.-J. (2021), ‘Constraining permanova and ldm to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data’, Microbiome **9**(1), 1–19.