Statistical Methods for the Analysis of Microbiome Data

Anna M. Plantinga

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Michael C. Wu, Chair

Ali Shojaie

Li Hsu

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Methods for the Analysis of Microbiome Data

Anna M. Plantinga

Chair of the Supervisory Committee:
Associate Member Michael C. Wu
Fred Hutchinson Cancer Research Center

The human microbiome plays a vital role in maintaining health, and imbalances in the microbiome are associated with a wide variety of diseases. Understanding whether and how the microbiome is associated with particular health conditions is a focus of many modern microbiome studies, with the hope that a deeper understanding of these associations may lead to more effective prevention and treatment regimens. However, how best to analyze data from microbiome profiling studies remains unclear. The high dimensionality, compositional nature, intrinsic biological structure, and limited availability of samples pose substantial statistical challenges. To face these challenges, we propose novel analytic approaches based on sparse penalized regression strategies and distance-based global association analysis.

Most distance-based methods for global microbiome association analysis are restricted to simple dichotomous or quantitative outcomes, but more complex outcomes are increasingly common in microbiome studies. In the first part of this dissertation, we introduce two distance-based methods for the analysis of entire microbial communities in modern microbiome studies. We develop a kernel machine regression-based score test for association between the microbiome and censored time-to-event outcomes. We then propose a novel longitudinal measure of dissimilarity that summarizes changes in the microbiome across time and compares these changes between subjects. Since this dissimilarity may be incorporated into any distance-based analysis framework, it is a highly flexible tool for applying a wide

variety of distance-based analyses in longitudinal studies.

Identification of associated taxa and detection of predictive microbial signatures are key to translation of microbiome studies. In the second part of this dissertation, we present two penalized regression methods for estimation and prediction with high-dimensional compositional data. Because phylogenetic similarity between bacteria often corresponds to shared functions, our first contribution is to incorporate phylogenetic structure into a penalized regression model for constrained data. We then propose a model that exploits phylogenetic structure to use partial information in the setting of differing feature sets between model-building and prediction datasets.

We evaluate the performance of these methods through extensive simulation studies and apply them to studies investigating the association of graft-versus-host disease or body mass index with the gut microbiome.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Michael Wu, for his mentorship over the past several years. His insights and enthusiasm — about biostatistics, the patterns and responsibilities of academic life, and how to recognize your particular gifts while also challenging yourself to grow — have been instrumental to my development as a biostatistician. His input has played a key role in this dissertation research. I am also grateful to my committee members Ali Shojaie, Li Hsu, and Johanna Lampe for their valuable feedback on this work.

I have benefited greatly from the guidance of many members of the University of Washington Department of Biostatistics. Sharon Browning, Brian Browning, Noah Simon, Katie Kerr, Scott Emerson, Mary Lou Thompson, and Barbara McKnight have all provided invaluable support and mentorship in research and teaching. I am deeply grateful to Gitana Garofalo for her advocacy for students and her support in both academic and personal difficulties. Many thanks also to the entering class of 2013, whose support and sense of humor made the years of coursework and qualifying exams fun.

DEDICATION

To my parents, who have loved, supported, and encouraged me every step of the way.

Chapter 1

INTRODUCTION

Humans exist in relationship with a diverse and extensive collection of microbes. A human body contains as many bacterial cells as human cells [112], and these bacteria collectively possess 150 times as many unique genes as are contained in the human genome [72, 103]. We use the term *microbiome* to refer to all of the microorganisms that inhabit the human body (the *microbiota*) and their gene content (the *metagenome*). The microbiome is of tremendous interest to basic science researchers and clinicians alike in the pursuit of a deeper understanding of human health. Microbial inhabitants of the human skin, intestinal tract, oral cavity, and several other body sites have been characterized in healthy individuals [51] and associated with a wide range of health conditions, including obesity [125], menopause symptoms [87], graft-versus-host disease [44], bacterial vaginosis [94], and type 2 diabetes [104]. The role of the microbiome extends to mediating disease treatment responses. For example, the gut microbiome is associated with efficacy of dietary interventions in irritable bowel disease [25] and is vital for the success of some cancer immunotherapies [107]. However, despite a keen scientific interest in gleaning insights from microbiome data, methods for statistical microbiome analysis have historically lagged behind the rapid advances in sequencing technology and the vastly expanding array of scientific and clinical microbiome studies, and statistical methods that incorporate unique features of microbiome data and accommodate modern study designs are needed. In this introduction, we first briefly review approaches to microbiome data generation, then provide an overview of several broad classes of statistical analyses for microbiome data. We end by outlining the aims of this dissertation.

1.1 Metagenomic Approaches to Microbiome Profiling

Prior to the development of next generation sequencing technologies, studies of the human microbiome required laborious cultivation of microbes from samples such as skin swabs, fecal samples, or endoscopic biopsies. Besides being time- and effort-intensive, culture-based microbiome studies are restricted to investigating taxa that can be grown in culture. Contrary to the long-held belief that only a tiny fraction of bacteria can be grown in the lab, emerging evidence demonstrates that most of the gut microbiota is culturable [14, 67]. However, different culturing procedures often reveal distinct microbial communities, and it remains unclear how to unify culture-based studies [70].

Metagenomic approaches to microbiome profiling provide a time-efficient, cost-effective, culture-free means to summarize the taxa present in a community. First, all of the DNA from a sample is extracted. DNA extraction is followed by either 16S rRNA sequencing or shotgun metagenomics [56]. The former proceeds by amplifying and sequencing the 16S rRNA gene, which is present in all bacterial species. This gene is attractive as a barcode because it contains highly conserved regions that enable PCR amplification, as well as nine "hypervariable regions" that are useful for taxonomic classification [17]. 16S rRNA reads are clustered into operational taxonomic units (OTUs) at a desired level of similarity, commonly 97% similarity [116]. The 16S sequences may also be used to build a *de novo* phylogenetic tree describing the evolutionary relationships between OTUs in the study or to place observed OTUs on a tree built from pre-existing reference databases [29, 34, 102]. Compared to shotgun metagenomics, this approach is inexpensive, simple to carry out, and accompanied by more extensive reference databases and analysis pipelines. In shotgun metagenomics, all genetic material present in the sample is sequenced, allowing much greater resolution of information about the functional potential of a particular microbial community, but sacrificing some ability to precisely identify taxa [64, 103, 127]. For this dissertation, we focus on 16S rRNA sequencing data summarized as a set of taxon counts for each individual, potentially accompanied by a phylogenetic tree summarizing the evolutionary similarity among

the taxa.

A limitation of 16S rRNA sequencing is that the total bacterial load in an environment (e.g., a human gut) is not ascertainable. The set of taxon counts therefore contains relative information about taxon abundance, but not absolute information. That is, samples with higher or lower proportions of, say, *E. coli* are distinguishable. However, from 16S data alone, inference about the true number (*absolute abundance*) of *E. coli* and other taxa is not possible, so we cannot determine whether the high proportion is due to an overgrowth of *E. coli* or a suppression of the rest of the microbiota. Alternative methods such as quantitative PCR can provide absolute abundance data for individual bacterial species or strains [32], but it is not feasible with current technologies to generate absolute abundances for the entire microbiome.

Because the total number of sequencing reads varies between samples, the taxon counts are not directly comparable across samples. Counts are therefore generally normalized to per-subject *relative abundances* (taxon proportions). Both absolute and relative abundances often play a role in bacterial community health [101]. Although absolute abundances contain more information than relative abundances, due to the current technological limitations for collecting absolute abundance data for more than a few taxa, most existing methods operate on relative abundances. Likewise, we henceforth consider only relative abundance data.

1.2 Statistical Approaches to Microbiome Data Analysis

We consider studies whose goal is to explore the association of the microbiome (as predictor) with a host phenotype such as disease or survival. Several features of microbiome data pose difficulties in statistical analysis. First, the normalization of microbiome data to proportions results in compositional data and induces correlation among taxa. This can lead to erroneous results using traditional statistical methodology if not accounted for [124]. Second, microbiome data are often statistically high-dimensional, with more taxa observed than subjects. Much effort has been put towards the development of statistical tools for high-dimensional but non-compositional data, and these tools can be leveraged and adapted for use with com-

positional data [15]. Finally, bacterial phylogeny is associated with presence and function, so leveraging phylogenetic information often provides higher power and accuracy across a range of statistical methods.

Scientific questions associated with microbiome analysis may be posed at the level of individual taxa, entire microbial communities, or both. The unique challenges and opportunities associated with microbiome data play a slightly different role depending on the scientific question and associated class of analyses. Often, broad scientific questions center around (1) each taxon's association with the outcome, (2) the role of diversity within a microbial community, (3) overall comparisons of distinct microbial communities, and (4) identification of taxa that are important or predictive in a particular setting.

First, individual taxon abundances may be compared between phenotypes using standard statistical methods or microbiome-specific approaches [81, 96]. Analysis at the taxon level does not need to explicitly account for the unit-sum constraint of compositional data provided that results are interpreted on a relative, not absolute, scale, although often data transformations that account for compositionality can improve the reliability and interpretability of results. Whether particular taxa are associated with an outcome is also clearly identified. These two factors make taxon-level analysis methodologically straightforward and scientifically interesting. However, due to the potentially large number of OTUs measured (hundreds to thousands), this route often requires stringent multiple testing corrections, resulting in lower power. In addition, the extensive network of taxon-taxon and taxon-host interactions may not be captured by analysis of individual bacteria.

Second, measures of within-sample diversity (α-diversity) may be compared across samples. Quantities such as the Chao1, Shannon, and Simpson indices summarize the estimated number of species present and the evenness with which the species are observed. In the gut, for instance, high α-diversity tends to indicate a healthy community, whereas low diversity has been linked to conditions such as inflammatory bowel disease; the reverse is true in the vagina, where a healthy community is dominated by *Lactobacillus* and high diversity is associated with bacterial vaginosis [51, 82, 115].

Distance-based methods constitute a third class of microbiome analyses. These use summaries of β-diversity (between-sample diversity) to compare overall taxonomic profiles between individuals. Distance-based analyses begin by computing pairwise dissimilarities between communities (samples), where the measures of dissimilarity are ecologically relevant and may incorporate phylogenetic structure. The matrix of pairwise dissimilarities may be summarized by its top principal coordinates for visualization, and permutation-based approaches or variance component score tests in a kernel machine regression framework are used for formal hypothesis testing. Distance-based analysis may provide power gains over taxon-level analysis by utilizing phylogenetic relationships among taxa, avoiding the multiple testing problem, and aggregating modest effects across multiple taxa [24].

Finally, machine learning methods have attracted significant recent interest for microbiome analysis. Statistically, microbiome data tends to be high-dimensional, with more taxa than subjects, and it is accompanied by extrinsic phylogenetic or functional structure. Scientifically, knowing that the microbiome is associated with a disease is rarely as relevant as knowing which bacterial taxa are driving the association, since the latter is operationalizable through intervention with probiotics, prebiotics, or antibiotics [100]. Also, microbiome-based predictive models are demonstrating utility for diagnostic and prognostic purposes [35, 59, 78, 141]. Statistical learning methods, particularly penalized regression models, are attractive because they are designed for high-dimensional settings, the choice of penalty function can impose structure, and penalties that induce sparsity both improve prediction and identify associated taxa.

1.3 Dissertation Aims

This dissertation focuses on distance-based and machine learning methods for microbiome data analysis. The vast majority of distance-based methods are designed for studies of the microbiome at a single time point with quantitative or categorical outcomes. However, modern microbiome studies encompass a much wider range of designs and outcome types. In Chapters 2 and 3, we develop methods for distance-based analysis in modern microbiome

studies. Chapter 2 presents MiRKAT-S, a kernel machine regression-based test for association between the microbiome and time-to-event outcomes. Because microbiome data are sparse and often have small sample sizes, the standard score statistic results in a highly conservative test. We develop a small-sample correction that preserves type 1 error and has high power. This work has been published in Plantinga et al. (2017). Chapter 3 introduces a novel longitudinal measure of dissimilarity that may be used with any existing β-diversity analysis, including ordination, permutation-based testing, and kernel machine regression-based score tests. The proposed dissimilarity, LUniFrac, summarizes changes in the microbiome between time points for each individual and compares these changes across individuals. In combination, these two methods extend distance-based analysis to microbiome studies generating time-to-event data (such as clinical trials) and longitudinal microbial profiles.

In Chapters 4 and 5, we develop structurally informed penalized regression methods for feature selection and prediction using the microbiome. Standard feature selection methods are not directly applicable in the setting of compositional data due to the negative correlation between proportions, and because evolutionarily similar taxa tend to be associated with similar outcomes, the incorporation of extrinsic phylogenetic groups is likely to improve selection and prediction accuracy. Chapter 4 presents the compositional sparse group lasso, which uses a linear log-contrast model to account for the compositionality of the data and applies ℓ_1 and ℓ_2 penalties to yield sparsity at two taxonomic levels of classification. In Chapter 5, we propose a multilevel linear log-contrast model that exploits the hierarchical compositionality of the microbiome. The model includes an aggregate group-level effect and modifies that effect for individual taxa using within-group proportions. To induce sparsity, ℓ_1 penalties are applied at both levels. An important advantage of the multilevel model is its flexibility for prediction in new data sets; the set of observed taxa differs between data sets, and this model exploits higher levels of taxonomic classification to use partial information for taxa that are present in the prediction set but not the training set. Chapter 6 concludes with a summary and suggestions for future research.

Chapter 2

DISTANCE-BASED MICROBIOME ANALYSIS FOR SURVIVAL OUTCOMES

Existing work on distance-based analysis of the microbiome has primarily focused on relatively simple dichotomous or quantitative outcomes such as disease status or biomarker levels. As microbiome profiling is increasingly incorporated into clinical studies, there is also considerable interest in the relationship between the microbiome and censored survival outcomes. However, standard dissimilarity based tests cannot accommodate censored survival times. In this chapter, we develop a new approach, MiRKAT-S, for community-level analysis of microbiome data with censored survival times. MiRKAT-S uses ecologically informative distance metrics, such as the UniFrac distances, to generate matrices of pairwise distances between individuals' taxonomic profiles. The distance matrices are transformed into kernel (similarity) matrices, which are used to compare similarity in the microbiome to similarity in survival times between individuals. Simulation studies demonstrate that our modified score statistic provides proper control of type 1 error and adequate power. We apply MiRKAT-S to examine the relationship between the gut microbiome and survival after allogeneic blood or bone marrow transplant. This work has been published in Plantinga et al. (2017) [99].

2.1 Background

Since taxonomic profiles are sparse and high-dimensional — hundreds to thousands of unique OTUs may be identified, many of which are present in only a few samples — comparisons on the level of individual OTUs may have low power. An alternative to OTU-level analysis is to compare the microbiome at the community level, i.e., to compare overall taxonomic profiles between individuals [21, 24, 57, 61, 99, 120, 125, 136, 144, 147]. This class of analyses is

often performed by computing pairwise distances between communities (samples), where the distance metrics are ecologically relevant and may incorporate phylogenetic structure. The matrix of pairwise distances is summarized by its top principal coordinates for visualization, and distance-based multivariate methods coupled with permutation are used to determine if dissimilarity is related to an outcome [6]. Distance-based analyses may provide power gains by utilizing phylogenetic information, avoiding the multiple testing problem, and aggregating modest effects across multiple taxa [24].

As an alternative to distance-based approaches that use permutation analysis, Zhao et al. [147] proposed the microbiome regression-based kernel association test (MiRKAT). MiRKAT uses a kernel machine framework with a variety of ecologically informative kernels to test for associations between the human microbiome and either continuous or binary outcomes. Intuitively, MiRKAT compares similarity in taxonomic profiles between communities (where similarity is measured via a kernel, which can be obtained by transforming relevant distance matrices) to similarity in outcome measures. P-values are obtained analytically using a variance-component score test. MiRKAT has the added advantages of flexible modeling of the relationship between the microbiome and outcome measures, natural incorporation of covariates, and efficient computation of p-values.

A limitation of existing community-level analysis approaches is that they cannot accommodate censored survival outcomes. However, such outcomes are of tremendous interest as microbiome profiling studies move into the clinical arena. For example, the lung microbiome has been related to progression of idiopathic pulmonary fibrosis [47] and the gut microbiome to overall survival after allogeneic blood and bone marrow transplant [55]. Additional OTU-level studies with survival outcomes have shown associations between the intestinal microbiome and development of allergic rhinitis [11] and atopic dermatitis [97] in children.

We therefore propose a test for association between the microbiome and censored survival outcomes (MiRKAT-S), accounting for covariates and potential confounders. We perform a distance-based analysis using the kernel machine Cox regression framework, encoding taxonomic profiles into kernel matrices via a transformation of distance metrics appropriate for

microbial communities. This allows the analysis to take into account phylogenetic information and other features specific to biological communities. To formally test the association between the microbiome, as encoded in the kernel matrix, and censored survival times, we use a variance component score test. However, when applied to microbial community profiles summarized by common kernels, the usual test statistic with p-values calculated by resampling procedures is highly conservative [16, 75]. We therefore implement a small sample correction that provides proper control of type I error while maintaining adequate power, and we calculate p-values analytically rather than by resampling. We demonstrate the performance of MiRKAT-S using real and simulated data summarized by a variety of kernels commonly used in microbial ecology.

This work represents the translation of existing methods in genetic studies with survival outcomes to applications in microbiome research. The first major contribution of this chapter is to allow survival outcomes in the kernel machine regression framework with kernels that appropriately encode microbiome data. Our small sample correction method provides proper control of Type I error and improved power when using microbiome-appropriate kernels, whereas the kernel machine regression based test as implemented for genetic studies has almost no power to detect relationships between microbial taxonomic profiles and survival. Secondly, the ability to perform the test using a variety of kernels provides robustness to the nature of the true association between the microbiome and survival. Therefore, although MiRKAT-S is technically similar to previous kernel machine regression methods, it enables microbiome analyses that are not possible using existing methods.

2.2 Microbiome Regression-Based Kernel Association Test for Survival

To associate the microbiome at the community level and censored survival times, we relate censored survival times to taxonomic profiles using a flexible non-parametric modeling framework. We assess significance via a variance component score test which acknowledges the modest sample sizes of most microbiome profiling studies. In this section, we first describe the modeling framework, followed by the testing strategy and technical advances necessary

to ensure proper control of type I error. Finally, we describe simulations encompassing a variety of true relationships between the microbiome and survival time.

2.2.1 Model Specification

Suppose that for each of n subjects, we observe the microbial taxonomic profile, encoded by a q-vector of OTU counts Z_i, and a p-vector of other covariates X_i. Let T_i be the survival time and C_i the censoring time for the ith subject. We observe the bivariate vector (Δ_i, U_i), where $U_i = \min(T_i, C_i)$ is the observed time and $\Delta_i = I(T_i \leq C_i)$ is the event indicator for subject i. We wish to test whether taxonomic profiles Z are associated with survival time, adjusting for covariates X.

The most commonly used model for censored survival times is the Cox proportional hazards model [30] due to its flexibility and relative robustness. Therefore, to relate T and (X, Z), we propose to use the kernel machine Cox proportional hazards model [16, 75],

$$\lambda(t; X, Z) = \lambda_0(t) \exp\left[X\beta + f(Z)\right] \tag{2.1}$$

where $\lambda_0(t)$ is the baseline hazard function. In the kernel machine regression framework, $f(\cdot)$ is generated by a positive definite kernel function $K(\cdot, \cdot)$, that is, $f(\cdot)$ lies in the reproducing kernel Hilbert space $\mathcal{H}_K$. Under the representer theorem [58], $f(Z_i) = \sum_{i'=1}^{n} \alpha_{i'} K(Z_i, Z_{i'})$ for some $\alpha_1, ..., \alpha_n$. Choosing different kernel functions $K(\cdot, \cdot)$ allows specification of a wide variety of models. For example, the kernel function $f(Z_i) = Z_i'\gamma$, corresponding to the linear kernel $K(Z_i, Z_{i'}) = Z_i Z_{i'}'$, is used to specify a linear model. Kernels are similarity matrices, so each element $K_{j,k} = K(Z_j, Z_k)$ represents the pairwise similarity between samples j and k. Because we use a score test, which depends only on the null model, any kernel will result in a valid test; however, the choice of kernel affects the power of the test.

To specify relevant models for microbial profiles, we use kernel functions that encode the similarity between the microbiome for two samples via a transformation of pairwise distance metrics. There are many commonly-used ecological distance and dissimilarity metrics, each with different features and strengths. For example, the UniFrac [79] and generalized UniFrac

[21] distances take into account the organization of OTUs into phylogenetic trees, thereby gaining power when clusters of taxa are associated with the outcome. Other dissimilarities, such as the Bray-Curtis dissimilarity [13], look at the presence and relative abundance of each OTU regardless of phylogenetic structure. These and other commonly-used distance metrics can be used to create distance matrices D, where each element d_{ij} is a pairwise distance between the taxonomic profiles of two samples. The distance matrices are then transformed to kernels, or similarity matrices, via

$$K = -\frac{1}{2}\left(I - \frac{11'}{n}\right) D^2 \left(I - \frac{11'}{n}\right)$$

as described in [147]. Here, I is the $n \times n$ identity matrix and 1 is an n-vector of ones. To ensure that K is positive semi-definite, we replace negative eigenvalues with zero. That is, we perform an eigenvalue decomposition $K = U\Lambda U$, where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$, and then reconstruct the kernel matrix using the nonnegative eigenvalues $\Lambda^* = \text{diag}(\max(\lambda_1, 0), ..., \max(\lambda_n, 0))$ so that $K = U\Lambda^* U$.

2.2.2 Score Test

Testing whether taxonomic profiles are associated with the outcome in the kernel machine Cox model corresponds to testing the hypothesis $H_0 : f(Z) = K\alpha = 0$. When the model is re-expressed using kernels as

$$\lambda(t; X, Z) = \lambda_0(t) \exp\left[X\beta + K\alpha\right]$$

where $K_{ij} = K(Z_i, Z_j)$, we can estimate (β, α) by maximizing the penalized log partial likelihood function

$$\log(PL) = \sum_{i=1}^{n} \int_0^{\infty} \log\left[\frac{e^{\beta' X_i + \alpha' K_i}}{\sum_{j=1}^{n} Y_j(s) e^{\beta' X_j + \alpha' K_j}}\right] dN_i(s) - \frac{c}{2}\alpha' K \alpha$$

where $N_i(s) = I(U_i \leq s)\Delta_i$, $c \geq 0$ is the penalty parameter, and $Y_j(s) = I(U_j \geq s)$ is an indicator that subject j is at risk at time s. An important relationship between kernel regression and linear mixed models has been described for non-censored outcomes [76]; a

similar relationship holds in the Cox model, as discussed in [16]. Therefore, solving the penalized log partial likelihood above is equivalent to fitting the frailty model

$$\lambda(t; X, Z) = \lambda_0(t) \exp[X\beta + h]$$

where $h = (h_1, ..., h_n)$ are random effects with mean 0 and variance τK. Then testing $H_0 : f(Z) = K\alpha = 0$ is equivalent to testing $H_0 : \tau = 0$, which can be accomplished using a variance-component score test. Since a score test only requires fitting the null model $\lambda(t; X, Z) = \lambda_0(t) \exp[X\beta]$, we do not need to estimate $f(Z)$, so the test is valid even if a nonoptimal kernel is used. However, choosing a kernel that accurately reflects the true relationship between the microbiome and survival time will provide higher power. Two factors determine how well the kernel reflects the true relationship: first, whether the abundance of the associated taxa matters (versus presence or absence), and second, whether the OTUs related to the outcome are clustered on a phylogenetic tree. For example, since the weighted UniFrac distance encodes both taxon abundance and phylogenetic information, a test based on the weighted UniFrac distance will have highest power when the true association is between the outcome and the abundance of a cluster of OTUs on a phylogenetic tree. Its power will be lower when the true association is with the abundance of unrelated OTUs (of similar frequencies) or with the presence or absence of a set of OTUs.

The variance component score statistic is

$$Q = \hat{M}' K \hat{M}$$

where $\hat{M} = (\hat{M}_1, ..., \hat{M}_n)$ is the vector of estimated martingale residuals under the null model, i.e., $\hat{M}_i = \Delta_i - \int_0^\infty Y_i(t) e^{\hat{\beta}' X_i} d\hat{\Lambda}_0(t)$ [16, 75]. Here $\hat{\Lambda}_0(u) = \sum_{i=1}^n \Delta_i I(U_i \leq u)/\hat{S}_0(U_i)$ is Breslow's estimator of the baseline hazard function $\Lambda_0(u) = \int_0^u \lambda_0(t)dt$ under the null model and $\hat{S}_0(t) = \sum_{i=1}^n Y_i(t) e^{\hat{\beta}' X_i}$ is the estimator for the baseline survival function.

Under the null hypothesis, Q asymptotically follows a mixture of chi-square distributions. The distribution has been derived for a linear kernel [20], but can be written in general form: by the central limit theorem,

$$K^{1/2} M \sim N(0, P_0^{1/2} K P_0^{1/2}) \tag{2.2}$$

where $P_0 = V - VX(X'VX)^{-1}X'V$ with $V = \text{diag}(\int_0^\infty Y_i(t)e^{\hat{\beta}'X_i}d\hat{\Lambda}_0(t) - w_i(\beta, t_i)^2)$ and $w_i(\beta, t) = e^{\hat{\beta}'X_i}/\hat{S}_0(t)$. Therefore,

$$Q \sim \sum_{i=1}^{n} \tilde{\lambda}_i \chi^2_{1,i}$$

where $(\tilde{\lambda}_1, ..., \tilde{\lambda}_n)$ are the eigenvalues of $P_0^{1/2}KP_0^{1/2}$ and $\chi^2_{1,i}$ are independent χ^2_1 random variables. Note that K need not be full rank for this distribution to hold, since $\lambda_i = 0$ for the terms associated with the singular components of K, so those components of the distribution will have weight zero [117].

Up to this point, we have assumed that there are no tied survival times. In practice, tied survival times are fairly common due to coarse time measurements resulting from specific visit schedules or study follow-up dates. We use the Efron approximation to accommodate tied survival times [37]. This approximation performs well even with relatively small sample sizes or a high proportion of ties [48].

2.2.3 Small Sample Correction

The test outlined above is highly conservative for modest sample sizes and complicated kernels, such as kernels commonly used for the microbiome (see Appendix A: Table A.1). Hence, we propose an approximate test using a modified score statistic that accounts for overdispersion. Analogous "small sample" corrections have been proposed for quantitative and binary traits [23]. Specifically, we propose the modified score statistic

$$Q^* = \frac{\hat{M}'K\hat{M}}{\hat{M}'\hat{M}}.$$

To derive the distribution of this statistic, we need the covariance of the residuals $\hat{M}$. We use a diagonal small-sample approximation to $Cov(\hat{M})$ that is motivated by the corresponding weighted linear model at convergence. This approximation is justified both in existing literature (e.g., [93, 95]) and through empirical evidence, namely the rapid convergence of the iteratively reweighted least squares (IRLS) algorithm using this weight matrix to the correct coefficients and the proper empirical type 1 error of our method.

Specifically, the fitted kernel machine Cox model (Equation 2.2.2) is equivalent to a weighted linear model at convergence with weight matrix estimated using an iteratively reweighted least squares (IRLS) algorithm, as described in Appendix A. We use a diagonal approximation for both the covariance of the residuals $\hat{M}$ and the weight matrix W for IRLS. Several versions of W have been used for weighted partial least squares in the literature (e.g., [68] and [93]); we use an intermediate version that is diagonal as in [93] and [95], but whose elements are defined by the negative Hessian with respect to β as in [68].

To express this mathematically, let $z = X\beta + W^{-1}\hat{M}$ be the working response. Again, although the covariance matrix of the residuals $\hat{M}$ is nondiagonal, we approximate $Cov(\hat{M})$ using a diagonal form proportional to the weight matrix W. Then by defining $z^* = W^{1/2}\tilde{y}$, $X^* = W^{1/2}X$, and $\epsilon^* = W^{1/2}\epsilon$, the weighted linear model can be written as

$$z^* = X^*\beta + \epsilon^*,\ \epsilon^* \sim N(0, \sigma^2 I)$$

with projection matrix $P_0^* = I - X^*(X^{*'}X^*)^{-1}X^{*'}$. At convergence, $\mathrm{Var}(\epsilon^*) = W^{1/2}\mathrm{Var}(\epsilon)W^{1/2}$ $= W^{1/2}\sigma^2 W^{-1}W^{1/2} = \sigma^2\mathbf{I}$.

Based on this, the distribution of Q^* satisfies

$$\begin{aligned} P(Q^* > q) &= P(\hat{M}'K\hat{M} - \hat{M}'q\hat{M} > 0) \\ &= P((P_0^*M)'(K - q\mathbf{I})(P_0^*M) > 0) \\ &= P(\epsilon' P_0^{1/2} P_0^*(K - q\mathbf{I})P_0^* P_0^{1/2}\epsilon > 0) \end{aligned} \tag{2.3}$$

where $\epsilon \sim N(0, \mathbf{I})$. The second equality uses $\hat{M} = P_0^*M$, as derived in Appendix A. The third equality uses the distribution of M that can be derived from Equation 2.2, that is,

$$M \sim N(0, K^{-1/2}P_0^{1/2}KP_0^{1/2}K^{-1/2}) \stackrel{d}{=} N(0, P_0).$$

Then under the null hypothesis,

$$Q^* \sim \sum_{i=1}^{n} \lambda_i^* \chi_{1,i}^2$$

where $(\lambda_1^*, ..., \lambda_n^*)$ are the eigenvalues of $P_0^{1/2}P_0^*(K - q\mathbf{I})P_0^*P_0^{1/2}$ and, as before, $\chi_{1,i}^2$ are independent χ_1^2 random variables. P-values can be calculated efficiently using Davies' exact

method [31]. For very small samples (e.g., $n \leq 50$), Davies p-values may be anticonservative and permutation p-values may be used instead.

2.3 Simulation Study

2.3.1 Simulation Scenarios

We carried out simulation studies in a range of settings to confirm that MiRKAT-S properly controls type I error and to assess its power using a variety of kernels. Microbiome OTU counts were generated using the same approach as [147]. Specifically, for each individual, we simulated OTU counts from a Dirichlet-multinomial distribution with dispersion parameters and proportions estimated from Charlson *et al.*'s real upper respiratory tract microbiome dataset, in which 856 OTUs were measured on each of 60 individuals [19]. The data for each simulated individual consists of 1000 total OTU counts distributed among the 856 OTUs of Charlson *et al.* We also simulated two covariates for each individual, X_{1i} and X_{2i}, from a standard Normal and a Bernoulli(0.5) distribution independently of taxonomic profiles. We considered sample sizes ranging from 25 to 500 individuals. For all simulation scenarios, we generated datasets with approximately 25% censoring. Four simulation settings were considered, varying (1) whether OTU abundance or the presence/absence of OTUs was associated with the outcome and (2) whether phylogenetically clustered or unclustered OTUs were associated with the outcome.

In setting 1, the abundances of OTUs in one cluster on a phylogenetic tree were associated with survival time. We partitioned all of the OTUs into 20 clusters using the partitioning-around-medoids algorithm based on the cophenetic distances of OTUs in the phylogenetic tree. The abundance of clusters ranged from 0.05% to 19.7% of all OTU reads. We selected an abundant cluster, containing 19.7% of all reads, to be associated with exponentially distributed survival times through the model

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X_i'\beta + \gamma \,\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right)\right)} \tag{2.4}$$

where γ is the true effect size for the cluster, λ is a scale parameter, $U_i \sim \text{Uniform}(0,1)$, $\mathcal{A}$ is the set of indices of OTUs in the selected cluster, and the "scale" function standardizes the total OTU abundance in the cluster to have mean 0 and standard deviation 1:

$$\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) = \frac{\sum_{j\in\mathcal{A}} Z_{ij} - \frac{1}{n}\sum_i \left(\sum_{j\in\mathcal{A}} Z_{ij}\right)}{SD_i\left(\sum_{j\in\mathcal{A}} Z_{ij}\right)}.$$

Censoring times were simulated independently as $C_i \sim \text{Exp}(\mu)$, and λ and μ are selected to give approximately 25% or approximately 50% censoring.

In setting 2, the ten most abundant OTUs overall, accounting for 31.5% of all OTU reads, were associated with survival time regardless of cluster membership. In this setting, we simulated survival times as

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X_i'\beta + \gamma\,\text{scale}\left(\sum_{j\in\mathcal{A}} \frac{Z_{i(j)}}{\bar{Z}_{(j)}}\right)\right)} \tag{2.5}$$

where $\bar{Z}_{(j)}$ is the average across samples of the counts for the jth OTU. This limits the ability of a single OTU to dominate the communal effect of the microbiome. Setting 2 is comparable to setting 1, since in both cases the abundance of common OTUs is associated with survival times, but it lacks setting 1's close phylogenetic relationship between associated OTUs.

In setting 3, the presence or absence of each OTU in a rare cluster, containing 0.9% of all reads, was associated with survival time. OTUs were clustered as in setting 1, but in this case, were associated with survival time via the model

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X_i'\beta + \gamma\,\text{scale}\left(\sum_{j\in\mathcal{A}} I(Z_{ij} > 0)\right)\right)} \tag{2.6}$$

Finally, in setting 4, the presence or absence of 40 randomly selected OTUs was associated with survival time. This mimics the size of an average cluster, since the mean number of OTUs assigned to a cluster was 42.8, with cluster sizes ranging from 3 to 118 OTUs. Since the majority of OTUs are rare, the overall number of OTU reads associated with the outcome is low in this setting. The model for T_i was the same as in setting 3. In both setting 4 and

setting 3, the presence or absence of rare OTUs is associated with survival times. However, setting 4 lacks setting 3's close phylogenetic relationship between associated OTUs.

In all simulation settings, we considered the weighted (K_w) and unweighted (K_u) UniFrac kernels, the Bray-Curtis kernel (K_{BC}), and the generalized UniFrac kernel with $\alpha = 0.5$ ($K_{0.5}$). These kernels are expected to have high power in different simulation settings. All of the UniFrac kernels take phylogenetic information into account. The unweighted UniFrac kernel does not account for OTU abundance, whereas the weighted UniFrac kernel does; the generalized UniFrac kernel is intermediate between weighted and unweighted. The Bray-Curtis kernel does not account for phylogenetic structure or overall abundance of an OTU, but does compare both presence/absence and relative abundance between samples of each OTU. Each kernel is expected to have highest power when its measure of distance (and therefore similarity) accurately reflects the true relationship between the microbiome and the outcome.

For each simulation setting, sample size n, and censoring proportion, and using each kernel, we applied the test described above to test for associations between OTU counts and survival time. We used 5,000 simulations with $\gamma = 0$ to estimate the empirical Type I error rate with a nominal significance level of 0.05 and estimated empirical power across a range of γ values using 1,000 simulated datasets.

We also compared MiRKAT-S to two alternative approaches sometimes used for community-level analysis. First, we considered OTU-level tests of all OTUs. For each of the 856 OTUs in the dataset, we ran a marginal Cox regression model. The minimum p-value from the 856 marginal models was compared to the null distribution to produce an overall p-value for any association of the microbiome with survival times. In practice, the null distribution would be generated for an individual study using permutation; however, in the interest of computational efficiency, we generated this distribution using the minimum p-values from 5000 simulations where survival times were not associated with the microbiome. Second, we performed principal coordinates analysis (PCoA) on a relevant distance matrix (see, e.g., [47]). Since it is not clear how to make PCoA plots with censored time-to-event outcomes,

we followed PCoA by Cox proportional hazards regression. Specifically, we generated the UniFrac and Bray-Curtis distance matrices as above, then included the top two principal coordinates as covariates in a Cox regression analysis. We tested the two principal coordinates jointly by using a chi-squared test to compare nested models with and without the microbiome-related predictors. Covariates X_1 and X_2 were included in all models exactly as in the MiRKAT-S simulations.

2.3.2 Simulation Results

Empirical type I error rates with 25% censoring are reported in Table 2.1. Equations 2.4, 2.5, and 2.6 are identical when $\gamma = 0$ (i.e., there is no true association between the microbiome and survival time), so settings 1-4 all have the same type I error. From the table, we see that MiRKAT-S is valid for all kernels and sample sizes of at least 100 individuals. For comparison, empirical estimates of type I error without the small sample correction are reported in Table A.1, demonstrating that the uncorrected test is highly conservative. Q-Q plots further clarify the behavior of p-values based on the corrected and uncorrected test statistics (Figure A.3). These plots show that p-values based on the corrected statistic do not deviate significantly from the theoretical distribution, whereas p-values based on the uncorrected statistic are far from the theoretical distribution. P-values based on the uncorrected statistic tend to be less extreme than they should be; p-values that are truly smaller than 0.2 tend to be overestimated, whereas p-values larger than 0.2 tend to be underestimated. For sample sizes smaller than 100, the size of MiRKAT-S is close to correct, though it may be slightly anticonservative. Empirical type I errors for small sample sizes are reported in Table 2.2. If the sample size is smaller than $n = 50$, it may be preferable to report p-values obtained using permutation.

To interpret the simulation results evaluating the power of the test, recall that two aspects of the relationship between the microbiome and survival are important for understanding which kernels should provide high power: the relationship between associated OTUs (whether or not they cluster on a phylogenetic tree) and the importance of taxon abundance (whether

Table 2.1: Empirical Type I errors for sample sizes n =100, 200, and 500 with approximately 25% censoring. Results are based on 5,000 simulated datasets. K_w, K_u, K_{BC}, and $K_{0.5}$ represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernel with $\alpha = 0.5$, respectively.

n	K_w	K_u	$K_{0.5}$	K_{BC}
100	0.0544	0.0540	0.0530	0.0542
200	0.0494	0.0480	0.0470	0.0462
500	0.0506	0.0478	0.0536	0.0442

Table 2.2: Empirical Type I errors for small sample sizes ($n < 100$) with approximately 25% censoring. Results are based on 5,000 simulated datasets and permutation p-values were obtained using 1000 permutations. K_w, K_u, $K_{0.5}$, and K_{BC} represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernel with $\alpha = 0.5$, respectively.

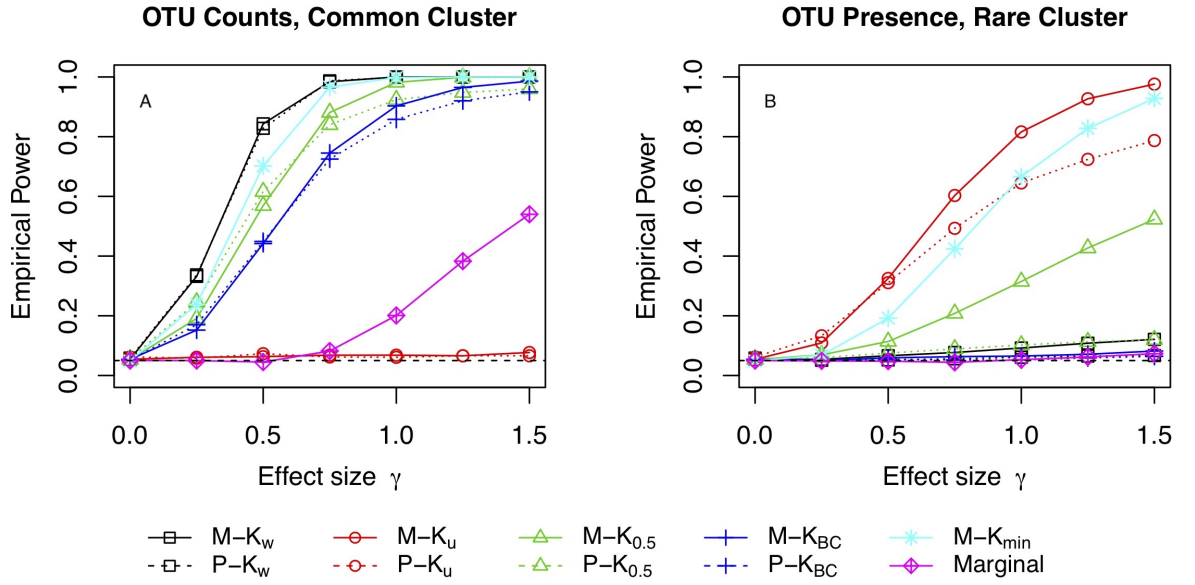
n	Method	K_w	K_u	$K_{0.5}$	K_{BC}
25	MiRKAT-S	0.054	0.055	0.055	0.056
	Permutation	0.046	0.048	0.046	0.049
50	MiRKAT-S	0.045	0.058	0.051	0.051
	Permutation	0.041	0.052	0.045	0.045
75	MiRKAT-S	0.054	0.058	0.051	0.053
	Permutation	0.051	0.053	0.048	0.049

OTU count or presence/absence matters). All of the UniFrac distances account for phylogeny, while the Bray-Curtis dissimilarity does not. The weighted UniFrac distance and Bray-Curtis dissimilarity both utilize taxon abundance (OTU counts), whereas the unweighted UniFrac distance only incorporates presence/absence of taxa, and the generalized UniFrac distance is intermediate between the weighted and unweighted UniFrac distances.

We first consider settings 1 and 3, in which a cluster of OTUs is associated with the outcome. When the OTU counts of an abundant cluster are associated with survival times (Figure 2.1A), the weighted UniFrac kernel and the generalized UniFrac kernel with $\alpha = 0.5$ provide the highest power, since the corresponding distance metrics take both abundance and phylogeny into consideration. Since the associated cluster is common, nearly all individuals have at least one read for each OTU in the cluster, so individuals cannot be distinguished based on OTU presence/absence. Therefore, in this setting, the unweighted UniFrac kernel has almost no power to detect the association. In contrast, when OTU presence/absence in a rare cluster is associated with survival time (Figure 2.1B), the unweighted UniFrac kernel has highest power, since this distance metric is based on presence and absence of OTUs. The weighted UniFrac kernel has very low power in this setting because OTU counts of a rare cluster do not vary much between individuals.

The power under settings 2 and 4, in which unclustered OTUs are associated with the outcome, is reported in Figure 2.2. When the OTU counts of the ten most common OTUs are associated with survival time (Panel A), the Bray-Curtis kernel has highest power, followed by the weighted UniFrac kernel and generalized UniFrac kernel with $\alpha = 0.5$. Since the Bray-Curtis dissimilarity metric does not incorporate phylogenetic information, this distance is designed for unclustered rather than clustered OTUs. However, since it takes abundance into account, the Bray-Curtis kernel performs better when OTU counts are associated with survival (e.g., Figure 2.1A) rather than OTU presence/absence (e.g., 2.1B) and when the associated cluster is at least moderately abundant. When the presence or absence of a random 40 OTUs were associated with survival time (Panel B), the unweighted UniFrac kernel was the only one with non-negligible power even at large effect sizes, but no kernel

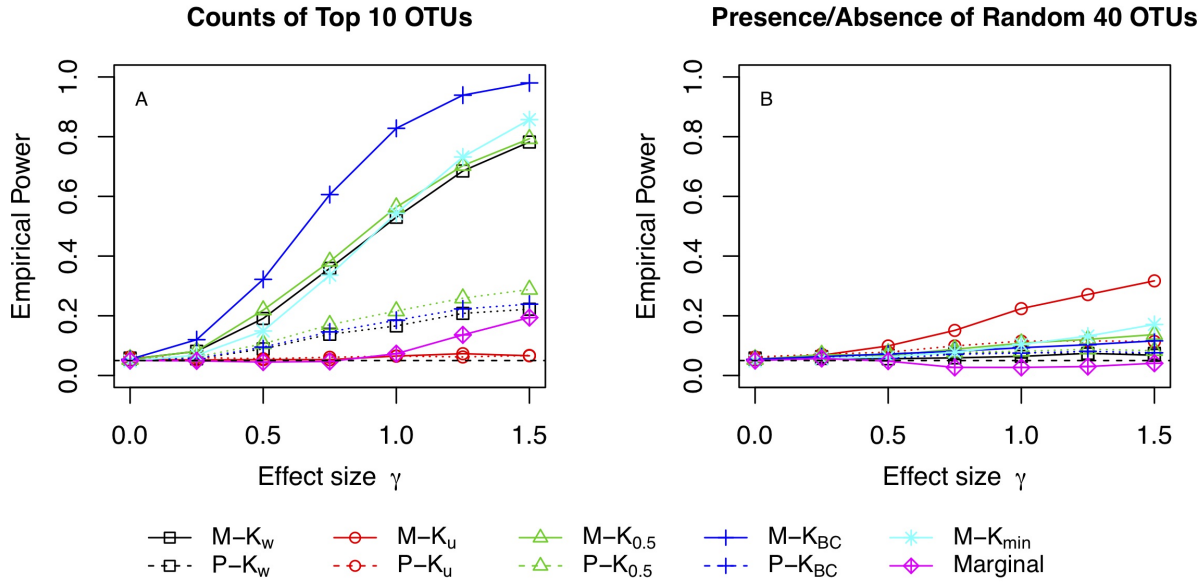
Figure 2.1: Empirical power for settings 1 (Panel A) and 3 (Panel B), using a sample size of n=100 and 25% censoring. M-K_w, M-K_u, M-$K_{0.5}$, and M-K_{BC} represent MiRKAT-S results for the weighted UniFrac kernel, unweighted UniFrac kernel, generalized UniFrac kernel with $\alpha = 0.5$, and Bray-Curtis kernel, respectively. P-K_w, P-K_u, P-$K_{0.5}$, and P-K_{BC} are the corresponding p-values from PCoA using the same set of kernels. M-$K_{\min}$ uses the minimum FDR-adjusted p-value from MiRKAT-S among all four kernels. "Marginal" refers to the permutation p-value summarizing OTU-level tests of all 856 OTUs.



had high power. The low power is due to the rarity of most randomly selected clusters and inability to gain power by utilizing phylogenetic information.

Kernel choice has a strong effect on the power of the test, and different kernels are optimal depending on the nature of the true relationship between the microbiome and survival time. In practice, a kernel representing relationships of particular scientific interest could be selected. For example, if a healthy microbiome at a certain body site has relatively few dominant taxa at high frequencies, and changes in the relative abundance of these taxa is hypothesized to be associated with the time to a disease outcome or death, choosing a kernel that accounts for taxon abundance will have the highest power to detect the hypothesized

Figure 2.2: Empirical power for settings 2 (Panel A) and 4 (Panel B), using a sample size of n=100 and 25% censoring. M-K_w, M-K_u, M-$K_{0.5}$, and M-K_{BC} represent MiRKAT-S results for the weighted UniFrac kernel, unweighted UniFrac kernel, generalized UniFrac kernel with $\alpha = 0.5$, and Bray-Curtis kernel, respectively. P-K_w, P-K_u, P-$K_{0.5}$, and P-K_{BC} are the corresponding p-values from PCoA using the same set of kernels. M-$K_{\min}$ uses the minimum FDR-adjusted p-value from MiRKAT-S among all four kernels. "Marginal" refers to the permutation p-value summarizing OTU-level tests of all 856 OTUs.



changes. If there is no specific hypothesized relationship, multiple kernels can be tested and then the resulting p-values adjusted for multiple comparisons. Testing the four kernels discussed here is a reasonable starting point, and the limited number of tests reduces the power loss due to adjusting for multiple comparisons. In these simulations, if the analysis is performed using the four kernels $K_u, K_{0.5}, K_w$, and K_{BC} and then the minimum p-value after an FDR adjustment is used for testing, the power does not quite match the best kernel, but is comparable to or better than the remaining three kernels (M-$K_{\min}$ in Figures 2.1 and 2.2).

We also compared the power of MiRKAT-S to two approaches used in current practice:

performing a marginal analysis for all OTUs, and including the top principal coordinates of the distance matrix as the covariates of interest in a regression model (dashed lines in Figures 2.1 and 2.2). In most simulation settings, MiRKAT-S has substantially better power than the marginal analysis or PCoA. In particular, the marginal analysis has power to detect an association between counts of OTUs in a cluster and survival times (Figure 2.1A), but virtually no power in any of the other settings. PCoA-based regression analysis performs similarly to MiRKAT-S for the best kernel when clustered OTUs are associated with survival times (Figure 2.1). However, for unclustered OTUs, PCoA has very low power for all kernels (Figure 2.2). Hence MiRKAT-S is more robust to kernel choice and true form of association than PCoA.

2.4 Application to Graft-Versus-Host Disease

Acute graft-versus-host disease (aGVHD) is a leading cause of death after allogeneic blood or bone marrow transplantation. There is a suspected relationship between the intestinal microbiome and aGVHD, but previous studies in mice and humans have yielded mixed results about the presence and nature of this relationship. Therefore, Jenq et al. recently studied the association of a particular bacterial species (intestinal *Blautia*) and of intestinal microbiome diversity indices with time to aGVHD onset, aGVHD-related mortality, and adverse outcomes unrelated to aGVHD [55].

In the original study, subjects were stratified into two cohorts depending on sequencing platform. The combined dataset used here results from resequencing of the first cohort of patients using the Illumina MiSeq platform; unfortunately, four patients did not have additional DNA available for MiSeq sequencing and were excluded from the analysis. Therefore 481 stool samples were available for 111 unique subjects, and for each sample, the Illumina MiSeq platform was used to sequence the V4-V5 region of the 16S rRNA gene. OTUs were generated as described in [55]. Briefly, mothur version 1.34 was used to compile and process sequence data [109], and quality filters were applied as in [108]. This procedure yielded counts for 2436 OTUs. As in [55], for each subject we only included the sample collected

closest to 12 days post-transplant in our analysis, and we excluded subjects for whom no samples were collected between 8 and 16 days post-transplant, so that 94 subjects were included in the final analysis. We used QIIME with default settings to align the sequences and generate a rooted phylogenetic tree. The 109 OTUs that failed to be placed on the tree were excluded, leaving 2327 OTUs. We applied MiRKAT-S using the unweighted and weighted UniFrac kernels, the generalized UniFrac kernel with $\alpha = 0.5$, and the Bray-Curtis kernel, adjusting for age and gender. The outcomes considered were overall survival and time to adverse event, where the adverse event includes stage 3 aGVHD, death, or relapse of the primary disease.

The results of applying MiRKAT-S to these data with and without the small sample correction are reported in Table 2. The association between overall survival and the microbiome is significant at $\alpha = 0.05$ using the unweighted UniFrac kernel K_u, generalized UniFrac kernel $K_{0.5}$, and Bray-Curtis kernel K_{BC}, but not using the weighted UniFrac kernel K_w (Table 2). For the composite endpoint, significant associations are seen using $K_{0.5}$ and K_w; however, this could be driven by the association with overall survival, which is included in the composite endpoint. The association remains significant after we adjust for multiple comparisons (multiple kernels) using either the false discovery rate method or the Bonferroni correction. The differences between the corrected and uncorrected p-values are fairly small. However, they are in the direction we would expect based on simulation results. In particular, we saw that low and high p-values are less frequent than would be expected for a null distribution of p-values (Figure A.3). This is consistent with seeing slightly higher p-values for the uncorrected statistic in these data, where the p-values based on the corrected statistic are fairly small.

To visualize the association between the gut microbiome and survival, we clustered individuals using Ward's agglomerative hierarchical clustering method [132] based on the generalized UniFrac distance with $\alpha = 0.5$. Ward's method is a generic clustering method that can be used for many data types. Generally speaking, the goal is to divide samples into clusters (groups) that tend to be similar in the ways that we care about; here, clusters should reflect

Table 2.3: Results of analysis of gut microbiome after allogeneic transplant. P-values from MiRKAT-S using the weighted (K_w) and unweighted (K_u) UniFrac kernels, the generalized UniFrac kernel with $\alpha = 0.5$ ($K_{0.5}$), and the Bray-Curtis kernel (K_{BC}). Adverse event (stage 3) refers to relapse, aGVHD stage 3, or death from any cause. "Corrected" indicates the p-values are based on the modified score statistic with proper type I error; "Uncorrected" indicates the p-values are based on the original score statistic. All analyses were adjusted for age and sex.

Outcome	Method	K_u	$K_{0.5}$	K_w	K_{BC}
Overall survival	Uncorrected	0.043	0.009	0.064	0.032
	Corrected	0.040	0.007	0.063	0.023
Adverse event (stage 3)	Uncorrected	0.077	0.007	0.034	0.089
	Corrected	0.076	0.006	0.033	0.082

similarity of taxonomic profiles. Operationally, Ward's method begins by assigning each sample to its own cluster and sequentially merges pairs of clusters that are most similar into larger clusters until all samples are merged into a single cluster. Which clusters to merge is decided by minimizing the increase in the sum of within-cluster squared distances (when Euclidean distances are used, this is the within-cluster variance). Through this process, a hierarchical tree is created. The tree can be cut at different levels to create the desired number of final clusters used for analysis. Although Euclidean distances are often used for Ward's method, other squared distances (in this case, ecologically relevant metrics such as the UniFrac distances) can be substituted while still using the same form of criterion and algorithm [91]. For our analysis, we used the generalized UniFrac distance to measure dissimilarity between individuals to ensure that clusters are similar with regard to the presence and abundance of taxa, accounting for phylogenetic relationships. We chose to cut the tree to create two clusters; a clear separation into clusters of sizes n=45 and n=49 can be seen in Figure 2.4A.

Kaplan-Meier curves for overall survival in the two clusters are shown in Figure 2.4B. However, the simple Cox regression p-value is not significant (p=0.09). Importantly, the similarity between individuals was measured the same way in both analyses, but MiRKAT-S

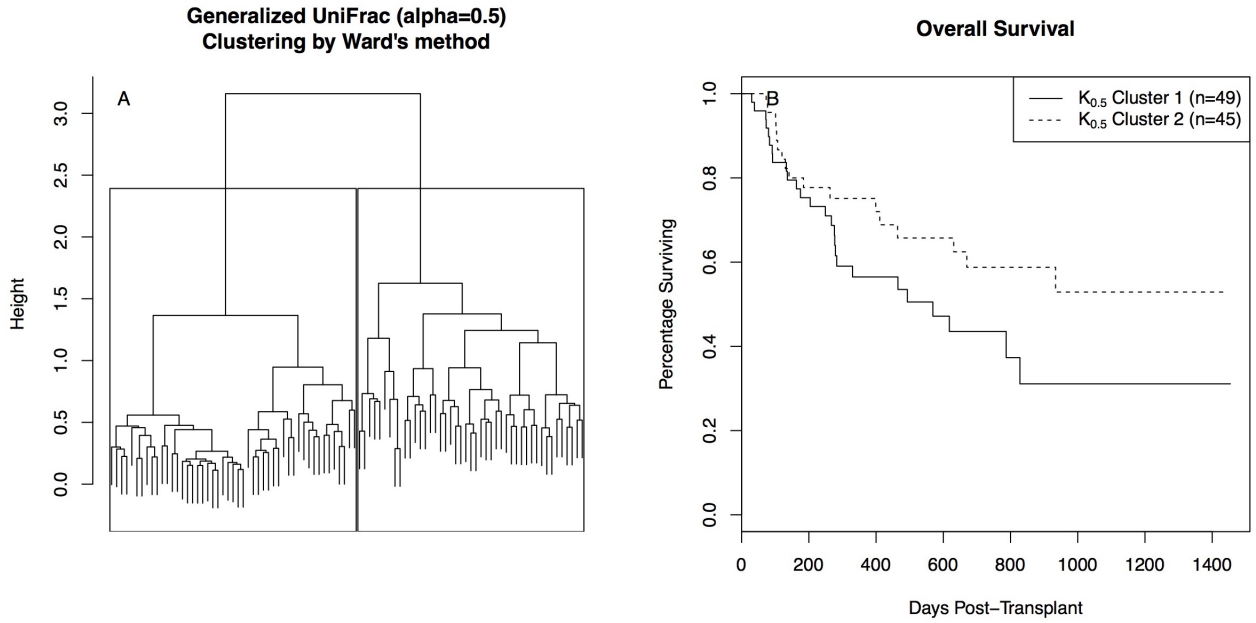


Figure 2.3: (A) Clustering of individuals using Ward's hierarchical clustering method, based on generalized UniFrac distances with $\alpha = 0.5$. (B) Kaplan-Meier curves for the two clusters of individuals with an outcome of overall survival.

yielded a highly significant p-value for the association of the microbiome with overall survival, whereas the analysis based on clustering individuals gave a nonsignificant result. Therefore, MiRKAT-S has higher power to detect this association than a simple clustering analysis based on the same distance metric.

The highly significant result using MiRKAT-S with $K_{0.5}$ may also provide information about the form of the association between the gut microbiome and survival post-transplant. The generalized UniFrac kernel incorporates phylogenetic information and represents a compromise between abundance and presence or absence of OTUs. Therefore, this kernel has highest power to detect relationships between taxonomic profiles and overall survival that occur through moderately rare clusters of OTUs or through a combination of common and rare clusters of OTUs. Accordingly, the high significance of MiRKAT-S using $K_{0.5}$ may in-

dicate that one of those settings holds: either moderately rare clusters of OTUs are driving the relationship between the microbiome and overall survival, or multiple clusters of OTUs, some of which are abundant and some of which are rare, are associated with overall survival. However, without further analysis we cannot determine which OTUs or clusters are associated with survival times in aGVHD patients.

2.5 Discussion

We propose MiRKAT-S for testing the association between the human microbiome and survival outcomes. In the kernel machine Cox regression framework, taxonomic profiles are modeled through a kernel function. This allows comparison of microbial community profiles using microbiome-specific distance metrics such as the UniFrac distances or Bray-Curtis dissimilarity. The kernel machine regression framework also allows linear (or more general parametric) adjustment for covariates and potential confounders. We test the significance of the association between the microbiome and survival times using a variance component score test, and we develop a small sample correction to account for the modest sample sizes and sparse, high-dimensional data that often result from microbiome studies. In contrast to existing methods that use resampling, p-values are obtained analytically using the Davies approximation.

Like other distance-based analyses, MiRKAT-S is limited to detecting the presence of an association between the microbiome and survival times. It cannot identify individual taxa that are associated with the outcome, and does not provide information about relationships among taxa within a microbial community. MiRKAT-S is therefore designed to be used when the question of interest is whether an entire microbial community is associated with the outcome. Alternative ways to answer this question include testing the association of each OTU individually with the outcome of interest or using a dimension reduction technique such as PCoA and testing the top few principal coordinates. Our simulation studies show that MiRKAT-S has power at least comparable to, and often substantially greater than, either of these methods for community level association testing. Community-level tests can be used

in combination with other methods that identify taxa of interest. These include marginal tests for particular OTUs of interest, identification of OTUs with high loadings from PCA or PCoA, or penalized regression methods that account for the structure and compositional nature of the data.

Our simulation results show that MiRKAT-S correctly controls type I error. However, under conditions of extreme censoring or very small sample sizes, the analytic p-values provided by MiRKAT-S may be slightly anti-conservative. In these cases, obtaining p-values by permutation may be preferable. Type I error is accurate regardless of the choice of kernel, but the power of the test depends heavily on how well the selected kernel encodes the true relationship between the microbiome and the outcome of interest. For example, when the abundance of an OTU or set of OTUs is related to the outcome, a kernel that encodes abundance information, such as the weighted UniFrac or Bray-Curtis kernel, will have higher power than a kernel that encodes only taxon presence or absence.

If there is no *a priori* hypothesis about which kernel will best represent relationships of scientific interest, the analysis can be performed using multiple kernels and an overall p-value can be obtained by permutation or adjustment for multiple comparisons. This analysis approach can provide information not only about the presence of a relationship, but also about its form, depending on the distance metrics considered and their relative power for different forms of the true association. That is, if the metric with the lowest p-value has highest power to detect associations with abundance of common clusters, that may be the form of the true association. Furthermore, weighted combinations of kernels could be used to simultaneously detect different types of shifts in the microbiome. Specific combinations or kernel weights could either be selected *a priori* or via a grid search, again using permutation to test overall significance. As the field of microbiome analysis matures and new distance metrics are proposed, our approach will continue to increase in power.

Chapter 3

TWO-STAGE UNIFRAC METRIC FOR LONGITUDINAL MICROBIOME ANALYSIS

The vast majority of distance-based methods, including MiRKAT-S, are only appropriate for independent samples; they cannot accommodate relationships present in longitudinal data. However, longitudinal studies are very attractive, since they allow comparison of a person's own "healthy" and "diseased" microbiome or association of microbiome changes with changes in disease state. This chapter develops a longitudinal UniFrac dissimilarity that summarizes within-individual shifts in microbiome composition, then compares these compositional shifts across individuals. This dissimilarity may be used in a wide variety of downstream analyses, including ordination analysis and distance-based hypothesis testing. Simulations show that tests based on the proposed dissimilarity retain appropriate type 1 error and high power. We apply the proposed dissimilarity to test the association between the gut microbiome and graft-versus-host disease.

3.1 Distance-Based Analysis for Longitudinal Microbiome Studies

Distance-based analysis can be modified to new settings, such as longitudinal studies, in two ways: by modifying the model used to relate pairwise distances to the outcome, or by modifying the distances themselves to accommodate additional information. In the previous chapter, we took the former approach to extend kernel machine regression-based analysis to survival outcomes. Along those lines, a linear mixed model framework has allowed the extension of formal distance-based association tests to longitudinal study designs for quantitative phenotypes [143]. However, similar longitudinal methods do not exist for more complex outcomes, such as time-to-event data, multivariate phenotypes, or even binary outcomes, and

specialized extensions would be required for each of these situations.

We instead consider modifying the distance metric. This approach has already proven valuable in the UniFrac family of distances: after the original proposal of (unweighted) UniFrac [79], which only considers taxon presence, differences in taxon abundance were incorporated into weighted UniFrac [80]. Variance-adjusted weighted UniFrac improved power by weighting differences in branch proportions by the corresponding variance [18], and generalized UniFrac moderates the weight placed on abundant or rare lineages [21]. Each of these adaptations increases the flexibility and information content in the measure of (dis)similarity used to compare microbial communities.

In this spirit, we propose the longitudinal UniFrac dissimilarity (LUniFrac) for microbiome studies with two time points per subject. LUniFrac computes a UniFrac-type distance between two subjects by comparing, instead of taxon abundances, a normalized measure of difference between the two time points for each subject. We develop both an unweighted and a generalized version of the dissimilarity. LUniFrac is a highly flexible tool for testing and ordination analysis, since it may be used in any existing testing framework or visualization procedure. Unlike a linear mixed model approach, LUniFrac explicitly summarizes and compares changes in the microbiome over time, permitting direct answers to the scientific questions often posed in longitudinal studies.

In the following sections, we present the unweighted and generalized LUniFrac dissimilarity metrics; perform simulation studies to verify that type 1 error control is maintained with LUniFrac and that it has power to detect true longitudinal associations; and apply the method to a dataset exploring the association of the gut microbiome with survival after allogeneic stem cell transplant.

3.2 Longitudinal β-Diversity Analysis

We propose two longitudinal measures of dissimilarity, analogous to the unweighted and generalized UniFrac distances. Both utilize a two-stage approach. In the first stage, the changes in taxon presence (unweighted) or abundance (generalized) for each subject are

summarized; in the second stage, these changes are compared across subjects, incorporating phylogenetic structure in much the same way as the other UniFrac distances. Figure 3.1 provides a visual representation of the procedure.

3.2.1 Unweighted LUniFrac

The original unweighted UniFrac metric sums the lengths of branches on a phylogenetic tree that are unshared between two microbial communities [79]. That is, if a taxon is present in one community but not the other, then the length of that taxon's branch of the tree contributes to the distance between the communities.

To extend this to two time points, we define change between time points for subject i and taxon j based on taxon presence or absence. Suppose we have measured OTU abundance for p taxa on n subjects at two time points, t_1 and t_2. Let p_k^{i,t_1} indicate the proportion of reads for subject i at time point t_1 that belong to taxon k. Then define

$$d_k^i(t_1, t_2) = I\left(p_k^{(i,t_2)} > 0\right) - I\left(p_k^{(i,t_1)} > 0\right) \in \{-1, 0, 1\}$$

for each subject $i = 1, \ldots, n$ and taxon $k = 1, \ldots, p$, where $I(\cdot)$ is the indicator function. Hence $d_k^i(t_1, t_2)$ is nonzero if and only if the taxon was present at exactly one of the measured time points for subject i; $d_k^i = 1$ if taxon k was present at time 2 but absent at time 1 (acquired between time points), and $d_k^i = -1$ if taxon k was present at time 1 but absent at time 2 (lost between time points). We will henceforth suppress the (t_1, t_2) notation and refer to these changes just as d_k^i.

The unweighted LUniFrac distance between subjects i and j is constructed based on d_k^i and d_k^j via

$$D_{ij} = \frac{\sum_{k=1}^{p} \frac{1}{2} b_k |d_k^i - d_k^j|}{\sum_{k=1}^{p} b_k}$$

so that D_{ij} summarizes the difference between subjects i and j in changes in taxon presence or absence, weighted by branch length and normalized to fall in [0,1]. In Appendix B we prove that this metric is a proper distance.

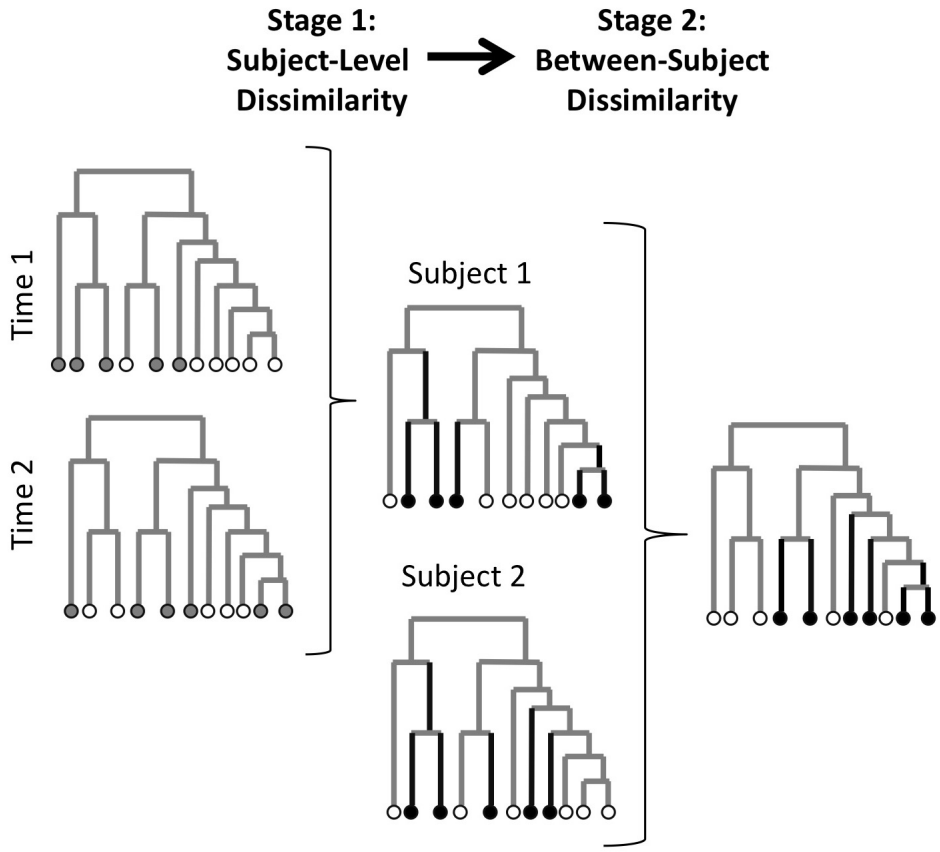


Figure 3.1: Schematic for calculation of unweighted LUniFrac metric. In Stage 1, gray circles indicate presence of the taxon and empty circles indicate absence. In Stage 2, black circles indicate changes from Time 1 to Time 2, and dark lines are phylogenetic distances that contribute to the LUniFrac dissimilarity. Pairwise dissimilarities are the differences in the taxa that changed from Time 1 to Time 2 between Subject 1 and Subject 2.

3.2.2 Generalized LUniFrac

Weighted UniFrac at a single time point extends the unweighted UniFrac distance by defining differences between communities in terms of differences in taxon proportion rather than taxon presence. Generalized UniFrac then weights each branch by a term that involves its overall abundance to avoid overweighting particularly rare or abundant lineages. The corresponding longitudinal dissimilarity, generalized LUniFrac, utilizes similar adjustments.

For generalized LUniFrac, we define change between times for subject i and taxon k based upon differences in abundance as

$$d_k^i(t_1, t_2) = \frac{p_k^{(i,t_2)} - p_k^{(i,t_1)}}{p_k^{(i,t_2)} + p_k^{(i,t_1)}} \in [-1, 1].$$

The sign indicates whether the taxon was more (+) or less (-) abundant at time 2 than at time 1, and the difference in proportion is normalized by overall taxon abundance. This normalization implies that the dissimilarity summarizes a quantity related to fold-changes in abundance, not absolute changes in abundance. The most extreme values are attained when a taxon is gained (+1) or lost (-1) entirely.

Using this measure of change, we then construct the weighted LUniFrac dissimilarity between subjects i and j via

$$D_{ij} = \frac{\sum_{k=1}^{p} b_k (\bar{p}_k^i + \bar{p}_k^j)^\alpha (\frac{1}{2}|d_k^i - d_k^j|)}{\sum_{k=1}^{p} b_k (\bar{p}_k^i + \bar{p}_k^j)^\alpha}$$

where $\bar{p}_k^i = \frac{1}{2}\left(p_k^{(i,t_1)} + p_k^{(i,t_2)}\right)$ is the average abundance of a particular taxon across times for subject i. Therefore D_{ij} summarizes the (normalized) difference between subjects of changes in taxon abundance across time, weighted by branch length and average taxon abundance. Any taxa that do not change in either community do not contribute to the distance; that is, we consider $b_k \times 0^\alpha \times \frac{0}{0}$ to be 0. Taken to the extreme, if neither community changes at all, i.e., $d_k^i = d_k^j = 0 \, \forall k$, we define $D_{ij} = 0$.

To better understand this measure of dissimilarity, notice that the term involving magnitude of change in abundance is $\frac{1}{2}|d_k^i - d_k^j|$, corresponding to the "weighting" term in weighted

UniFrac. This takes its largest value, 1, if $d_k^i = 1$ and $d_k^j = -1$ or vice versa, which happens if taxon k is both gained in subject i and lost in subject j or vice versa. It takes its smallest value, 0, if taxon k's abundance changes equally in the two subjects. This is already normalized to the absolute abundance of a taxon through the definition of d_k^i. For example, in this term, a change from a relative abundance of 0.2 to 0.1 results in exactly the same d_k^i as a change from a relative abundance of 0.4 to 0.2. The "generalization" (similar to generalized UniFrac) comes into play in the weighting of branch lengths by average abundance, $(\bar{p}_k^i + \bar{p}_k^j)^\alpha$. This term does involve absolute proportions, so in our toy example, it would weight a taxon with average abundance of 0.3 differently than a taxon with average abundance of 0.15. The parameter α controls the weight on abundant branches, so that larger α places higher weight on the contribution of common taxa, whereas small α places similar weight on common and rare taxa.

In Appendix B we show by counterexample that this measure of dissimilarity is not guaranteed to satisfy the triangle inequality, so it is not a true distance. Its performance in practice is not compromised by this observation; the generalized UniFrac measure also is not guaranteed to satisfy the triangle inequality, and yet it is one of the most widely-used dissimilarities in microbiome analysis.

3.2.3 Ordination Analysis and Testing

β-diversity metrics have wide-ranging utility in microbiome data analysis. Unweighted and generalized LUniFrac may be utilized in any analysis where a measure of β-diversity is required. Four main uses for measures of β-diversity are data visualization, creation of low-dimensional representations of the microbiome for incorporation in downstream models, classification and clustering, and global hypothesis testing. We outline each of these uses below.

In ordination analysis, high-dimensional data are mapped into a low-dimensional space, often using just two or three dimensions, so that similar observations lie near each other in the low-dimensional space and dissimilar observations lie far from each other. Several well-

known ordination methods are principal components analysis (PCA), multidimensional scaling (MDS) and non-metric multidimensional scaling (NMDS), and principal coordinates analysis (PCoA), although many others exist. Once the data are represented in low-dimensional space, observations may be plotted along these axes to visualize dissimilarity in the microbiome across several groups [38, 139]. The low-dimensional representation may also be included as a covariate or outcome measure in further analyses [90, 104].

For classification and clustering, the goal is again to explore relationships among samples, in this case by linking progressively more closely related samples. Clustering algorithms include hierarchical clustering, in which similarity between observations may be represented on a dendrogram, and discrete clustering methods such as K-means clustering or partitioning around medoids (PAM), which result in unstructured subgroups of samples. These types of methods have been used, for example, in relation to the idea of distinct "enterotypes" in the gut microbiome [8, 63]. Although recently enterotypes have been increasingly viewed along a gradient rather than as discrete categories [54, 60], discrete categorization remains a useful descriptive tool.

Finally, global hypothesis testing may be carried out by testing whether β-diversity differs across values of the outcome of interest. The category of distance-based multivariate analysis includes, among others, permutation-based methods such as PERMANOVA [6] and kernel machine regression-based association tests [24, 99, 136, 144, 147]. All of these formally test whether individuals with more similar outcomes also tend to have more similar microbiomes (as measured by β-diversity).

Because all of these analyses rely on β-diversity, LUniFrac provides a straightforward means of extending each of these analyses to explore change in the microbiome across time.

3.3 Simulation Studies

3.3.1 Simulation Methods

Simulations were performed to verify that use of LUniFrac dissimilarities preserves type 1 error control in existing kernel machine regression (KMR)-based global association tests and compare the power of the unweighted and generalized LUniFrac kernel with different choices of α across association settings. We generated OTU counts at the first time point from a Dirichlet-multinomial distribution with parameters estimated from real respiratory-tract data [19], as previously described [99, 147]. The dataset includes 856 OTUs, for which we generated 1000 observations (read counts) per sample. The second time point for each subject was generated by perturbation of the OTU counts from the first time point.

The perturbation weights were generated as $w_k \sim 1.25 \times Poisson(3) \times \min(0.54, Expo(3))$, and the perturbed OTU counts were rounded to the nearest integer. The Poisson component allows exact zeros in the weight vector, corresponding to taxa that disappear from a community entirely. The truncated Exponential component perturbs abundance on a continuous scale. The parameters of these distributions were chosen by trial and error with several objectives. First, we chose a distribution with mean 1 and variance 1 so that the overall mean, variance, and covariance of OTU counts in each cluster does not change. This prevents any effect of differential sequencing depth on the results; in practice, a similar end is often attained by rarefaction or other normalization procedures. The empirical moments of the distribution of w are $E[w] = 1.003$ and $Var[w] = 1.001$. Second, we required a relatively small proportion of exact zero weights, since relatively few taxa are expected to disappear entirely. For example, in the gut microbiome data considered in Section 3.4, the median proportion of taxa changing from nonzero to zero values between the two time points is 5.5%. While this proportion will vary between body sites and clinical settings, a relatively low proportion is reasonable. Empirically, the chosen distribution yields a 5.0% probability of the weight being exactly zero. Finally, we conservatively chose a modest maximum change in abundance, although in practice the fold-change in taxon proportion between time points

may be very large for some taxa. The IQR of the weights in the selected distribution is (0.29, 1.47) and the maximum is 11. Zero counts can never change to nonzero counts based on this perturbation scheme; allowing taxon introduction would improve power of the unweighted LUniFrac metric, but it does not affect type 1 error and is unlikely to substantially impact the performance of the generalized LUniFrac metric.

Quantitative, dichotomous, and time-to-event outcomes were simulated using changes in OTU presence or abundance rather than presence or abundance at one time point. OTUs were assigned to each of 20 clusters using the Partitioning Around Medoids (PAM) algorithm. A moderately common cluster of OTUs, comprising 11.8% of all reads, was selected to be associated with the outcome in simulations where magnitude of change in proportion matters, and a moderately rare cluster of OTUs, comprising 3% of all reads, was associated with the outcome in simulations considering presence of changes. Continuous outcomes were simulated as in [147] under the model

$$y = 0.5X_{1i} + 0.5X_{2i} + \gamma \,\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$ and $Z_{ij} = d^i_j(t_1, t_2)$ is the normalized change in taxon proportion or presence. The active set, $\mathcal{A}$, denotes the set of OTUs in the associated cluster. $X_{1i} \sim N(0,1)$ and $X_{2i} \sim Bernoulli(0.5)$ are time-invariant covariates, and the scale() function standardizes the total OTU abundance in the associated cluster to have mean 0 and variance 1. Similarly, binary outcomes were simulated under the model

$$\text{logit}\left(E(y_i|X_i, Z_i)\right) = 0.5X_{1i} + 0.5X_{2i} + \gamma \,\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right)$$

Finally, as in [99], survival times were simulated via

$$T_i = \frac{-\log(U_i)}{\exp\left(X_i'\beta + \gamma \,\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right)\right)}$$

where $U_i \sim Uniform(0,1)$, and censoring times were generated independently from the microbiome to yield approximately 25% censoring. For type 1 error simulations, we set $\gamma = 0$.

To examine the power of unweighted LUniFrac, we also considered a case-control design, in which the outcome was very rare but was strongly associated with the presence of a rare cluster of taxa. We generated data for excess individuals and preferentially sampled an equal number of cases and controls. The model for simulating case-control outcomes was

$$\text{logit}(P(y_i|Z_i)) = -4.59 + 4.59 \times \gamma \,\text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right)$$

where $Z_{ij} = d_j^i(t_1, t_2)$ as defined for unweighted LUniFrac. Using this model, without the presence of at least one member of the rare cluster, the probability of being a case was 1%. With at least one member of the cluster present, the probability of being a case increased to 50% with an effect size of $\gamma = 1$, and larger effect size γ or presence of more members of a cluster corresponded to larger probabilities.

3.3.2 Size and Power of KMR-Based Tests

We first verify that the KMR-based tests for longitudinal, dichotomous, and time-to-event outcomes have appropriate size using kernels computed from LUniFrac dissimilarities [99, 147]. Based on analysis using the R package MiRKAT with LUniFrac dissimilarities and no true association, type 1 error is indeed controlled at or near the nominal level of $\alpha = 0.05$ (Table 3.1).

We present power results for continuous and binary outcomes; results for time-to-event outcomes are qualitatively similar. In the setting where the magnitude of changes in taxon abundance in a moderately common cluster is associated with the outcome (left hand plots of Figure 3.2), the generalized LUniFrac dissimilarity with $\alpha = 0.25$ has highest power. The parameter α controls how much weight is placed on high-abundance branches, with higher α placing the most weight on branches with high abundance. This is consistent with our simulation procedure, in which changes in a moderately abundant taxon are associated with survival time. When changes in presence or absence of a rare cluster is associated with a continuous or binary outcome, the unweighted LUniFrac kernel has by far the highest power of any single kernel, whereas the generalized UniFrac kernel has negligible power regardless of

Table 3.1: Empirical size for each outcome type based on 2000 simulations, with n = 50, 100, 200, or 500 and nominal level $\alpha = 0.05$.

Outcome Type	n	K_{UW}	$K_{0.5}$	K_W	K_{omni}
Continuous	50	0.05	0.049	0.048	0.048
	100	0.051	0.049	0.054	0.051
	200	0.053	0.046	0.048	0.05
Binary	50	0.05	0.051	0.048	0.049
	100	0.045	0.049	0.045	0.047
	200	0.049	0.049	0.045	0.048
Time-to-Event	50	0.057	0.056	0.053	-
	100	0.054	0.053	0.055	-
	200	0.050	0.051	0.048	-

the choice of α. As is the case at single time points, the omnibus test has power close to that of the best-performing kernel. The omnibus test therefore provides an attractive alternative to choosing a single kernel, since in real-data settings, the true form of the association between the longitudinal microbiome and the outcome is rarely known in advance.

3.4 Application to Graft-Versus-Host Disease

We now return to acute graft-versus-host disease (aGVHD), which occurs in 30-70% of allogeneic blood or bone marrow transplant patients and is a leading cause of death following transplant. As mentioned in the previous chapter, Jenq et al. [55] recently studied the association of diversity of the gut microbiome and abundance of the genus *Blautia* with time to severe aGVHD, aGVHD-related mortality, and overall mortality. Although longitudinal microbiome profiles were collected, the original study focused on the post-transplant microbiome. However, using the pre-transplant microbiome as a baseline may clarify whether shifts from each subject's baseline microbiome are associated with adverse post-transplant outcomes.

The data were processed as described in the previous chapter. We excluded any samples with fewer than 500 reads. To assess the sensitivity of this analysis to read depth variability,

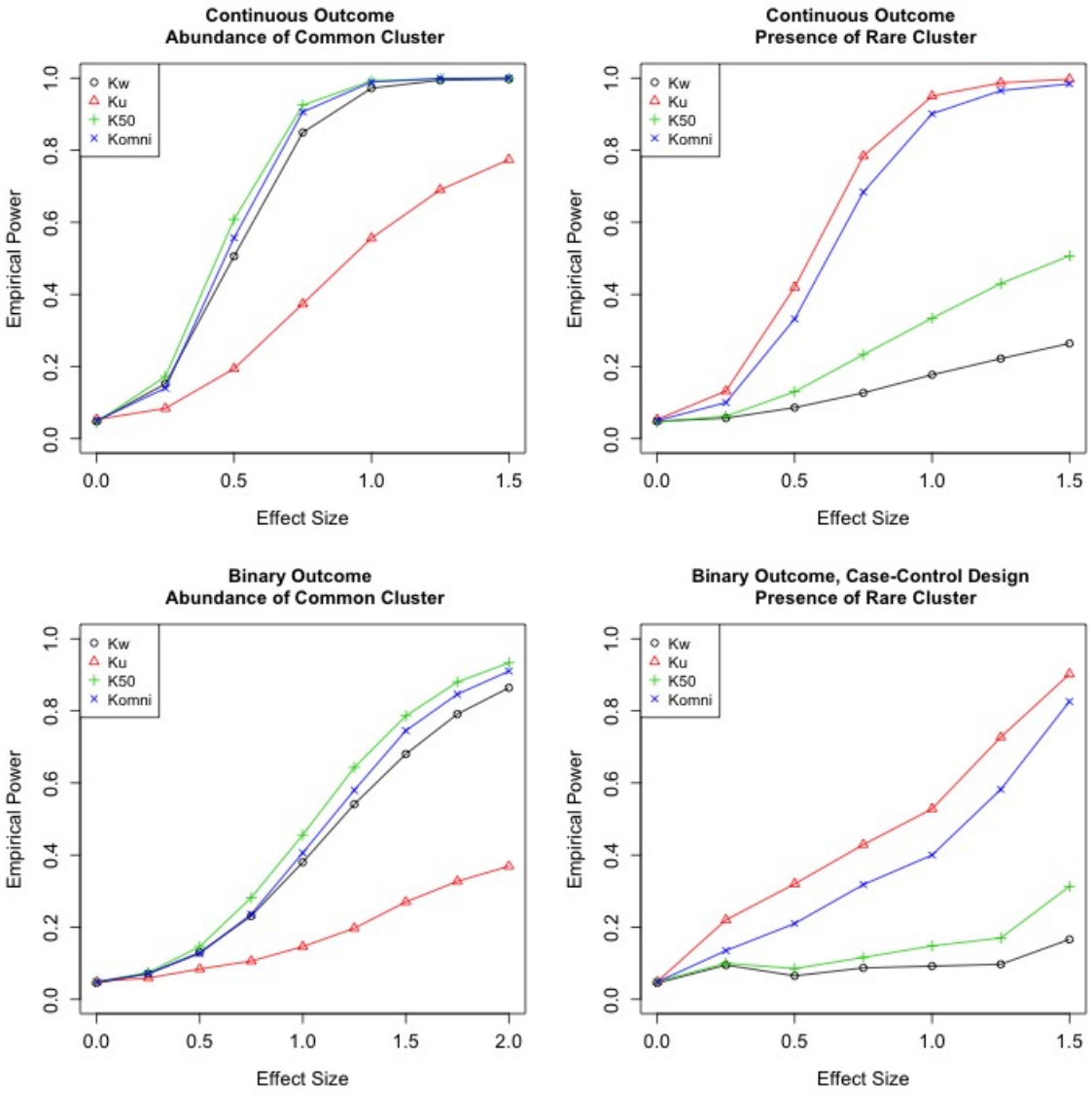


Figure 3.2: Empirical power based on 2000 simulated datasets with n = 200. Kw indicates the weighted LUniFrac kernel; Ku indicates unweighted; and K25, K50, and K75 indicate the generalized LUniFrac kernel with $\alpha = 0.25$, 0.5, and 0.75 respectively. Komni is the omnibus test, testing significance based upon all of the other kernels by permutation.

we also considered excluding samples with fewer than 1000 reads and rarefying to 500 or 1000 reads; results for all of the sensitivity analyses were similar to the primary results. For each subject, we included the last sample taken pre-transplant as our first time point (range: 1-14 days pre-transplant) and the sample collected closest to day 12 post-transplant as our second time point (range: 5-26 days post-transplant). Subjects missing a pre-transplant sample were excluded, so that a total of 97 subjects were included in the analysis. All analyses were adjusted for age and sex.

We first visualize the data using principal coordinates analysis (PCoA) on the LUniFrac dissimilarities. We adjust for age and sex by defining

$$K = -\frac{1}{2}(I - H)D^2(I - H)$$

using $H_1 = [1, \text{Sex}, \text{Age}]$ and $H = \text{scale}\,(H_1 H_1^\top)$, where I is the $n \times n$ identity matrix and the scale() function centers and scales the columns of H to have mean 0 and standard deviation 1. D is the dissimilarity matrix, and D^2 is the elementwise square; in this case, we choose the generalized LUniFrac dissimilarity with $\alpha = 0.5$ for D. The matrix K is then used for PCoA. In the PCoA plot (Figure 3.3), we indicate overall time on study by color and death due to any cause by point shape. The color gradient outward from the upper left region of the plot demonstrates that subjects with more similar observed times tend to have more similar changes in their microbiomes over that time period; there is no clear separation between subjects who were observed to die and those who were censored.

Using MiRKAT-S with LUniFrac dissimilarities to test the association between longitudinal microbiome and adverse outcomes after allogeneic transplant, we find very strong evidence for an association between changes in the gut microbiome and overall survival (Table 3.2). These results are consistent with the Chapter 2 results, which included only the post-transplant samples, but are substantially more significant and are significant across a broader range of kernel choices. This suggests that looking at global change from baseline microbiome provides more power than simply looking at the post-transplant microbiome. Additionally, we find an association between the gut microbiome and the aggregate event

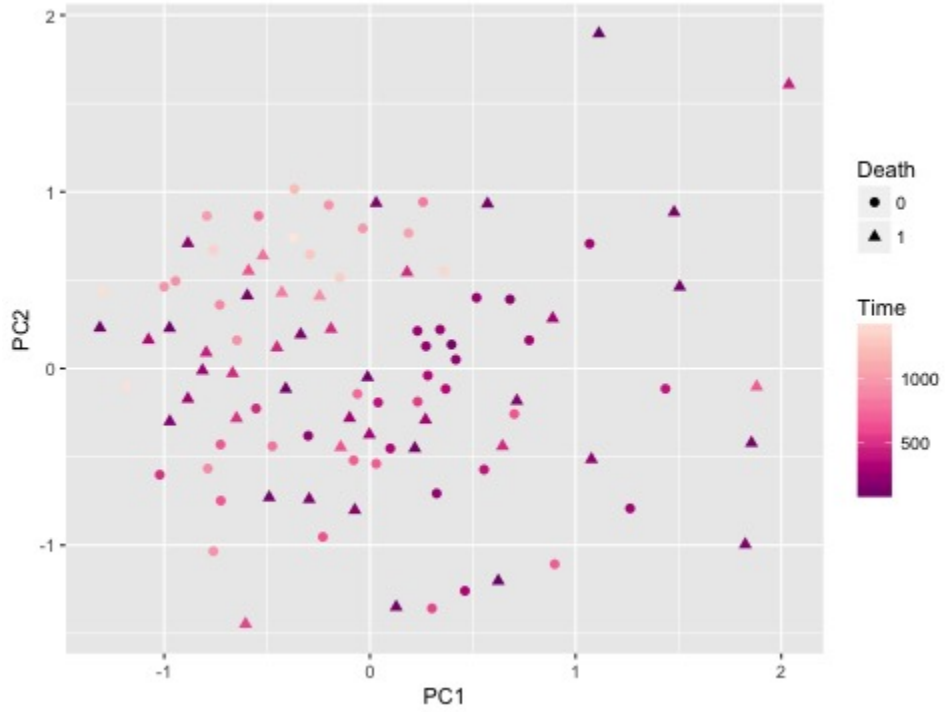


Figure 3.3: Principal coordinates plot using generalized LUniFrac dissimilarity with $\alpha = 0.5$. Color indicates observed time on the study (in days), and shape indicates whether death was observed.

aGVHD stage 2, relapse, or death using $K_{0.25}$ and $K_{0.5}$. Although this association could be driven either by overall survival or by other adverse events, the aggregate outcome is of clinical interest in GVHD studies of its own accord [50]. We also performed marginal Cox regression on the normalized changes between time points for each taxon, and after false discovery rate adjustment, no taxa were marginally significant. This demonstrates that studying changes between the pre- and post-transplant gut microbiome globally can improve power to detect clinically relevant associations, beyond that available via marginal analysis or global analysis at a single time point.

Table 3.2: P-values from MiRKAT-S using unweighted and generalized LUniFrac kernels, indicated by K_{UW}, K_α for $\alpha = 0.25, 0.5, 0.75$, and $K_W = K_1$. "Adverse event" refers to relapse, aGVHD of the specified stage, or death from any cause. Analysis was adjusted for age and sex.

Outcome	K_{UW}	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	K_W
Overall survival	0.642	0.002	0.001	0.002	0.004
Adverse event (stage 2)	0.068	0.014	0.044	0.086	0.152
Adverse event (stage 3)	0.623	0.066	0.087	0.100	0.113

3.5 Discussion

We have developed a measure of dissimilarity, LUniFrac, that compares changes in OTU presence or abundance across time between different individuals. LUniFrac dissimilarities may be used in any existing distance-based analyses, including testing in the kernel machine regression framework, PERMANOVA, visualization and analysis using PCoA, and other distance-based global microbiome tests. LUniFrac dissimilarities provide high power to detect associations between changes in the microbiome and clinical, biological, or environmental outcomes. Because it can be used in existing testing frameworks, tests based on LUniFrac can be very fast, adjust for relevant covariates, and accommodate a variety of outcome types.

As in generalized UniFrac and other families of distances, the power of unweighted and generalized LUniFrac depends on the true form of the association between changes in the microbiome and the relevant outcome. Unweighted LUniFrac summarizes the acquisition or loss of OTUs between an individual's two time points, then compares these dichotomized changes across individuals. It therefore tends to have high power when the presence of rare taxa (and in particular, the loss or acquisition of rare taxa) drives the association. The generalized LUniFrac distances have highest power to detect large changes in abundance of taxa relative to their average abundance. In each case, in contrast to the UniFrac family of distances at a single time point, the driving factor in LUniFrac is not the absolute taxon

abundance or presence, but rather how much those values change.

Although LUniFrac explicitly accommodates exactly two time points, more than two time points per subject may also be considered using the existing framework for omnibus tests. Specifically, LUniFrac dissimilarity matrices may be computed between each pair of time points, and then a permutation-based omnibus test allows the overall test of all kernels (and therefore all pairs of time points) under consideration. These tests are available for quantitative and dichotomous outcomes through MiRKAT or OMIAT [61, 147] and for time-to-event outcomes through MiRKAT-S or OMiSA [62]. Hence, LUniFrac is a powerful and flexible tool for balanced longitudinal studies of the microbiome.

Chapter 4

COMPOSITIONAL SPARSE GROUP LASSO

The previous two chapters address distance-based analyses, which have high power to detect associations but cannot identify which taxa are responsible for an association. However, identification of associated taxa is vital for the interpretation and translation of microbiome research. In this chapter, we turn to the use of machine learning methods in microbiome analysis, which can select associated taxa, estimate the form of the association, and make predictions based on the microbiome. With careful model and penalty choice, these methods are also able to account for the data's compositional nature and extrinsic structure based on phylogenetic or functional relationships among taxa.

We propose a constrained sparse group lasso, where the constraint accounts for the compositionality of the data and the group structure represents prior knowledge about taxon phylogeny or function. We formulate this problem as a convex optimization problem and use an alternating direction method of multipliers algorithm with accelerated generalized gradient descent for optimization. The proposed estimator satisfies model selection consistency and estimation consistency under fairly standard conditions. We evaluate the method's performance in simulation studies and apply it to examine the relationship between the gut microbiome and body mass index.

4.1 Feature Selection with Compositional Covariates

The number of taxa observed in a microbiome study is often similar to or greater than the number of subjects. In the setting of generic high-dimensional data, regularized regression methods such as the lasso and the elastic net are a common approach to estimation and feature selection [121, 148]. These methods have also been employed to understand

the association between the microbiome and outcomes including inflammatory markers after allogeneic stem cell transplantation [133], pharmaceutical weight-loss treatments [129], and chronic functional constipation in children [33]. However, the compositionality of 16S microbiome data induces negative correlation between taxon proportions, which traditional regularized regression methods are not designed to accommodate. Coefficient interpretation is also challenging because no single taxon may change in relative abundance without corresponding changes in other taxa.

Several regularized regression methods have been developed specifically for microbiome data. Kernel-penalized regression is a generalized form of ridge regression that accounts for data compositionality with the centered log-ratio (CLR) transformation, then applies a penalty based on taxon covariance to account for structure within the CLR-transformed data [105]. An alternative kernel-based method is variance component lasso selection (VC-lasso) [142]. This approach uses a linear mixed model with variance components corresponding to taxa at different levels of phylogenetic grouping, and taxon selection results from applying an ℓ_1 penalty to the variance components. Employing a more traditional linear regression framework, Lin et al. (2014) apply an ℓ_1 penalty to a linear log-contrast model, thereby accounting for compositionality, inducing sparsity, and yielding interpretable coefficient estimates [74]. This approach demonstrably outperforms the ordinary lasso applied to compositional data in terms of prediction error as well as selection and estimation accuracy. We will focus on this modeling framework.

However, the features in microbiome datasets (taxa) also have external structure due to functional or phylogenetic similarity that may be incorporated to improve model performance. Bacteria that are related phylogenetically often share functional capabilities, exhibit comparable activity levels, and are found in similar environments, even at the level of high taxonomic ranks such as phyla [9, 89, 98]. Functional profiling of the microbiome is also of increasing interest in microbiome analysis, particularly with the observation that taxonomically distinct microbial communities may have similar distributions of metabolic pathways [51]. Statistical approaches that incorporate external phylogenetic or functional structure

often demonstrate better performance (in terms of power, predictive accuracy, or cluster discrimination, for example) than corresponding unstructured methods [22, 79, 131].

In order to incorporate this external grouping structure, we propose applying a sparse group lasso penalty to the linear log-contrast model to form the compositional sparse group lasso (CSGL). This penalty selects zero and nonzero groups of coefficients, then further selects among features in nonzero groups [114]. In the setting of microbiome data, the ℓ_2 penalty of CSGL sets coefficients for entire groups of taxa (such as phyla or classes) to zero, whereas the ℓ_1 penalty promotes sparsity within groups (for example, at the genus level). All taxon effects are estimated at the lower taxonomic level. Because of the high degree of variability in the microbiome due to environmental factors and the resulting potential for confounding in microbiome association studies, we additionally allow a set of unpenalized features.

The remainder of the chapter is structured as follows. In Section 4.2, we describe the penalized linear log-contrast model and resulting convex optimization problem, which we solve using an alternating direction method of multipliers algorithm. In Section 4.3, we derive conditions under which the CSGL problem has an optimal solution that satisfies selection and estimation consistency. We explore the performance of CSGL under a variety of simulation scenarios in Section 4.4, considering settings in which the associated taxa follow the extrinsic grouping structure and in which the assumed structure does not describe the true association. In Section 4.5 we apply CSGL to a dataset relating BMI to the gut microbiome. Section 4.6 concludes with a discussion.

4.2 Methods

4.2.1 Compositional Sparse Group Lasso

The compositional sparse group lasso (CSGL) is based on the linear log-contrast model, originally proposed by Aitchison and Bacon-Shone (1984) for mixture data [3]. Suppose for each of n subjects we observe a quantitative outcome y_i and relative abundances of p

taxa $(X_{i1}, \ldots, X_{ip})$. Each row of X is therefore made up of components of a composition $(\sum_j X_{ij} = 1)$ lying in the $(p-1)$-dimensional positive simplex S^{p-1}. Due to this constraint, one element X_{ij} cannot be altered without a corresponding change in another element, hampering model identifiability and coefficient interpretation in traditional regression models.

To address this issue, Aitchison and Bacon-Shone propose applying a log-ratio transformation, in which one component is selected as the reference component and the remaining components are divided by the reference [3]. Taking the log of these ratios of proportions yields an unconstrained $(p-1)$-dimensional covariate matrix for which standard linear regression may be used. Expressed formally, the linear log-contrast model is

$$y = Z^p \beta_{\setminus p} + \epsilon \tag{4.1}$$

where $\epsilon \sim N(0, \sigma^2)$, Z^p is the matrix of log-ratios with reference component p and elements $Z^p_{ij} = \log(X_{ij}/X_{ip})$, and $\beta_{\setminus p} = (\beta_1, \ldots, \beta_{p-1})^\top$. However, the choice of reference component in this model has a strong effect on the fitted coefficients. Setting $\beta_p = -\sum_{i=1}^{p-1} \beta_i$ yields the symmetric form of the model,

$$y = Z\beta + \epsilon \quad \text{s.t.} \ \sum_{j=1}^{p} \beta_j = 0, \tag{4.2}$$

where $Z_{ij} = \log(X_{ij})$ and $\beta = (\beta_1, \ldots, \beta_p)^\top$. Model (4.2) is equivalent to (4.1), but no explicit choice of reference component is required, and coefficient estimates are invariant under permutation of the components of the composition.

Due to the high dimensionality of microbiome data, in which many more taxa may be observed than the number of samples, Lin *et al.* [74] apply an ℓ_1 penalty to Model (4.2). The addition of the penalty term makes the model identifiable even in high-dimensional settings, and the choice of an ℓ_1 penalty induces sparsity among estimated taxon effects. This model has lower prediction error and better estimation accuracy (assessed by absolute or mean squared deviation from true coefficients) than applying standard lasso regression to the unconstrained model.

However, prior knowledge about group membership of taxa may be exploited to further improve estimation and selection accuracy. We assume the p taxa are divided into q groups of size $p_1, \ldots, p_q$ according to higher-level phylogenetic classification (e.g., phylum or class) or major functional contribution. In order to account for the grouping structure, in the spirit of the sparse group lasso [114], we apply both ℓ_1 and ℓ_2 penalties to (4.2). The constrained convex optimization problem is therefore

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(||y - Z\beta||_2^2 + \lambda(1-\theta) \sum_{j=1}^{q} \sqrt{p_j} ||\beta^{(j)}||_2 + \lambda\theta ||\beta||_1 \right) \quad \text{s.t.} \sum_{k=1}^{p} \beta_k = 0, \quad (4.3)$$

where λ is the regularization parameter, θ defines the convex combination between ℓ_1 and ℓ_2 penalties, and $\beta^{(j)}$ refers to the coefficients in group j. The weights $\sqrt{p_j}$ balance the penalty between groups of different sizes. The resulting problem retains the benefits of interpretation and identifiability obtained under the compositional lasso, but has the potential for better selection and prediction accuracy due to the incorporation of additional structural information.

Because the linear log-contrast model requires log-transforming the taxon counts, zeros may not be included in the count matrix. Several methods have been proposed for handling zeros [134], including adding a small constant known as a pseudocount [81] or using a Bayesian formulation with a Dirichlet prior [83]. Each of these choices in some settings will affect analysis results, but comparing normalization methods is outside the scope of this paper. For the data application in Section 4.5, we take the common approach of adding a pseudocount of 0.5, the maximum rounding error.

4.2.2 Optimization Algorithm

We solve Problem (4.3) using an alternating direction method of multipliers (ADMM) algorithm. We first express the constraint as an indicator function $g_{\mathcal{C}}(\beta)$ such that $g_{\mathcal{C}}(\beta) = 0$ if $\beta \in \mathcal{C}$ and $g_{\mathcal{C}}(\beta) = \infty$ otherwise, where $\mathcal{C} = \{\beta \in \mathbb{R}^p : \sum_{j=1}^{p} \beta_j = 0\}$. Introducing α such

that $\alpha = \beta$, the problem may be rewritten as

$$\underset{\alpha,\beta}{\operatorname{argmin}} \left(\frac{1}{2n}||y - Z\beta||_2^2 + \lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}||\beta^{(j)}||_2 + \lambda\theta||\beta||_1 + g_{\mathcal{C}}(\alpha) \right) \text{ s.t. } \alpha = \beta \quad (4.4)$$

so that the augmented Lagrangian is

$$\mathcal{L}_\mu(\alpha,\beta,\xi) = \frac{1}{2n}||y-Z\beta||_2^2+\lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}||\beta^{(j)}||_2+\lambda\theta\|\beta\|_1+g_{\mathcal{C}}(\alpha)+\frac{\mu}{2}||\beta-\alpha+\xi||_2^2-\frac{\mu}{2}||\xi||_2^2 \quad (4.5)$$

where ξ is the scaled Lagrange multiplier and μ is the augmented Lagrangian parameter, which also serves as the step size for the dual variable updates. Our ADMM algorithm alternates between minimizing $\mathcal{L}_\mu(\cdot)$ with respect to β, α, and ξ, using accelerated generalized gradient descent with step-size selection for the β updates. The algorithm is outlined below; further details are given in Section 1 of the online Supplementary Materials.

1. Initialize α^0 and β^0 with 0 or a warm start, $\xi^0 = 0$, $\mu > 0$, and iteration index $k = 0$.

2. Update β via $\beta^{k+1} \leftarrow \operatorname{argmin}_\beta \mathcal{L}_\mu(\beta; \alpha^k, \xi^k)$

3. Update α via $\alpha^{k+1} \leftarrow P_{\mathcal{C}}(\beta^{k+1} + \xi^k)$, where $P_{\mathcal{C}}$ is the projection onto $\mathcal{C}$

4. Update ξ via $\xi^{k+1} \leftarrow \xi^k + \beta^{k+1} - \alpha^{k+1}$

5. $k \leftarrow k + 1$; repeat (2)-(4) until convergence.

Under Section 3.2.1 of Boyd et al. [12], this algorithm is guaranteed residual convergence ($A\alpha+B\beta \to 0$), objective convergence (the objective function approaches the optimal value), and dual variable convergence. Assumptions required for this result are verified in Appendix C.

4.2.3 Tuning Parameter Selection

As with the non-compositional sparse group lasso, the choice of tuning parameter is important for the performance of CSGL. The regularization parameter λ may be selected by

cross-validation or by the generalized information criterion (GIC) proposed by Fan *et al.* [39] and adopted by Lin *et al.* [74], defined by $GIC(\lambda) = \log \hat{\sigma}^2_\lambda + (s_\lambda - 1)\frac{\log\log n}{n}\log(\max(p,n))$ where $\hat{\sigma}^2_\lambda = ||y - Z\hat{\beta}_\lambda||^2_2/n$, $\hat{\beta}_\lambda$ are the coefficient estimates with regularization parameter λ, and s_λ is the number of nonzero coefficients in $\hat{\beta}_\lambda$. The parameter $\theta \in [0,1]$ is selected in advance to provide the desired trade-off between group-level and feature-level sparsity, where $\theta = 0$ is a compositional group lasso without any within-group sparsity and $\theta = 1$ is the compositional lasso. Depending on the strength of the grouping variable, we find that $\theta = 0.05$, $\theta = 0.5$, and $\theta = 0.95$ tend to perform well. Alternatively, both tuning parameters may be chosen by cross-validation or GIC over a grid of (λ, θ), although this is computationally more demanding. The step-size for the Lagrangian updates, μ, does not affect algorithm convergence or prediction error, and we use $\mu = 1$ throughout.

4.3 Theoretical Properties

We demonstrate that with non-vanishing positive probability, Problem (4.6) has an optimal solution that satisfies model selection consistency and bounded ℓ_∞ loss. Proofs of all results presented in this section are included in Appendix C. The proof of the main result follows the form of Lin *et al.* [74], but requires additional optimality and irrepresentable conditions due to the addition of the ℓ_2 penalty term.

We assume without loss of generality that the last component, Z_p, is chosen as the reference; that it belongs to the last group (q); that $\hat{\beta}_p \neq 0$; and that the columns of Z are normalized so that $\max_{k\in(1,\dots,p)}\|Z_k\|_2 \leq \sqrt{n}$. The form of the CSGL problem that is most convenient for the statement and proof of the main result is

$$\hat{\beta}_{\setminus p} = \underset{\beta_{\setminus p}}{\operatorname{argmin}} \left(\frac{1}{2n}\|y - Z^p\beta_{\setminus p}\|_2^2 + \lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}\|(D\beta_{\setminus p})^{(j)}\|_2 + \lambda\theta\|D\beta_{\setminus p}\|_1 \right) \tag{4.6}$$

where $D = \begin{bmatrix}\mathbf{I}_{p-1} & -\mathbf{1}_{p-1}\end{bmatrix}^\top \in \mathbb{R}^{p\times(p-1)}$ so that $\beta_p = -\sum_{i=1}^p \beta_i$. We denote by $(D\beta_{\setminus p})^{(j)}$ the elements of the vector $(\beta_1, \dots, \beta_{p-1}, \beta_p = -1^\top_{p-1}\beta_{\setminus p})$ corresponding to group j.

Let $\beta_k^{(j)}$ indicate the kth feature in the jth group. We define $\mathcal{G} = \{1 \leq j \leq q : \beta^{(j)} \neq 0\}$ to be the set of groups with at least one truly nonzero element β. $\mathcal{S}_j = \{1 \leq k \leq p_j : \beta_k^{(j)} \neq 0\}$

indicates nonzero features in group j only, and $\mathcal{S}_{\mathcal{G}} = \bigcup_{j \in \mathcal{G}} \mathcal{S}_j$ indicates nonzero features in any group. Hats indicate the corresponding sets for $\hat{\beta}$, and superscript p indicates the corresponding set excluding reference element p.

In the spirit of the Irrepresentable Condition for the lasso and group lasso, a key condition for our result bounds the correlation between zero and nonzero features. We assume here that $\theta \notin \{0, 1\}$; very similar conditions may be derived in either of those cases, without the term involving θ on the right hand side of the inequality. $\Sigma^p = \frac{1}{n}(Z^p)^\top Z^p$ denotes the sample covariance matrix for the log-ratio transformed data with reference element p. Submatrices are indicated by subscripts, so for example, $\Sigma^p_{\mathcal{S}^c_{\mathcal{G}} \mathcal{S}^p_{\mathcal{G}}}$ is the submatrix of the sample covariance that includes entries (i, j) such that $i \in \mathcal{S}^c_{\mathcal{G}}, j \in \mathcal{S}^p_{\mathcal{G}}$.

Condition 1. *There exists some $\eta \in (0, 1]$ such that*

$$\sqrt{\kappa} \left\| \Sigma^p_{\mathcal{S}^c \mathcal{S}^p_{\mathcal{G}}} (\Sigma^p_{\mathcal{S}^p_{\mathcal{G}} \mathcal{S}^p_{\mathcal{G}}})^{-1} \left\{ sgn(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - sgn(\hat{\beta}_p) \mathbf{1}_{s-1} \right\} + sgn(\beta_p) \mathbf{1}_{(r+t)} \right\|_\infty \leq \frac{(1-\eta)}{2} \cdot \frac{\min(\theta, 1-\theta)}{\max(\theta, 1-\theta)}$$

$$\sqrt{\kappa} \left\| \Sigma^p_{\mathcal{S}^c \mathcal{S}^p_{\mathcal{G}}} (\Sigma^p_{\mathcal{S}^p_{\mathcal{G}} \mathcal{S}^p_{\mathcal{G}}})^{-1} \left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}} \hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_q \hat{\beta}_p \mathbf{1}_{s-1} \right) + b_q \beta_p \mathbf{1}_{(r+t)} \right\|_\infty \leq \frac{(1-\eta)}{2} \cdot \frac{\min(\theta, 1-\theta)}{\max(\theta, 1-\theta)}$$

for fixed θ and $\kappa = \max_j \sqrt{p_j}$, where the set $\mathcal{S}^c = \mathcal{S}^c_{\mathcal{G}^c} \bigcup \mathcal{S}^c_{\mathcal{G}}$ includes all features with truly zero coefficients, regardless of whether the entire group is zero or nonzero; s is the number of nonzero β; and r and t are the number of zero βs in nonzero and zero groups, respectively.

The first equation in Condition 1 matches Condition 1 in [74], except for the additional factor of $\sqrt{\kappa}$ and the form involving θ on the right hand side of the inequality. The second equation is similar to the group-lasso consistency conditions in Bach *et al.* [10]; in our notation, their condition is

$$\max_{j \in \mathcal{G}^c} d_j^{-1} \left\| \Sigma_{\mathcal{S}^c_j \mathcal{S}} \Sigma^{-1}_{\mathcal{S}\mathcal{S}} \hat{B}_{\mathcal{S}_{\mathcal{G}}} \hat{\beta}_{\mathcal{S}_{\mathcal{G}}} \right\|_2 \leq 1$$

where d_j are fixed weights, specialized in our case to the square root of group size κ. The difference in weights is because our condition is on the ℓ_∞ norm whereas the condition in [10] is on the ℓ_2 norm. The extra terms $\hat{b}_q \hat{\beta}_p \mathbf{1}_{s-1}$ and $b_q \beta_p \mathbf{1}_{r+t}$ result from the compositionality constraint. These terms are analogous to the $\text{sgn}(\beta_p)$ terms in the first equation.

We now turn to our main result. Under fairly standard conditions on the minimum signal size and regularization parameter, along with the Irrepresentable Conditions, with positive nonvanishing probability an optimal solution exists that satisfies model selection consistency and bounded ℓ_∞ loss.

Theorem 1. *Suppose Condition 1 holds, the minimum signal size satisfies* $\beta_{min} > \frac{\psi}{2}((2\kappa+1)\lambda_1+3\lambda_2)$*, and the regularization parameter satisfies* $\lambda = c_1\sigma\sqrt{\kappa}\{\log(2p)/n\}^{1/2}/(\eta\ \min(\theta, 1-\theta))$ *for the selected* θ *and some constant* $c_1 > 2\sqrt{2}$*. Then with probability at least* $1 - 2p\exp\{-n\min(\theta,1-\theta)^2\eta^2\lambda^2/(8\kappa\sigma^2)\}$*, Problem (4.3) has an optimal solution* $\hat{\beta}$ *that satisfies:*

1. *Sign consistency:* $sgn(\hat{\beta}) = sgn(\beta)$*, and*

2. *Bounded* ℓ_∞ *loss:* $\|\hat{\beta}_{\mathcal{S}_\mathcal{G}} - \beta_{\mathcal{S}_\mathcal{G}}\|_\infty \leq \frac{\psi}{2}\{(2\kappa+1)\lambda_1 + 3\lambda_2\}$.

These constants differ from those for the compositional lasso only by constants. Hence, as in [74], if $\log p = o(n)$, Theorem 1 implies that the model selection and estimation consistency of the proposed estimator hold asymptotically, with convergence rate $\|\hat{\beta}_{\mathcal{S}_\mathcal{G}} - \beta_{\mathcal{S}_\mathcal{G}}\|_\infty = O_p\left[\{(\log p)/n\}^{1/2}\right]$ if the smallest possible λ is chosen.

4.4 Simulation Study

4.4.1 Simulation Settings

In this section, the performance of CSGL is evaluated relative to the compositional lasso [74], kernel-penalized regression using Aitchison's variation matrix (non-phylogenetic scenarios) or patristic distances (phylogenetic scenario) between features [105], and non-compositional analogues to each of these methods. We consider scenarios with either non-phylogenetic or phylogenetic grouping and correlation structures.

In the non-phylogenetic scenarios, our simulation strategy is similar to that of [74]. We generate a matrix of taxon proportions from a logistic normal distribution [1] by first simulating $W_{n\times p} \sim N_p(\theta, \Sigma)$, then transforming W to a compositional matrix X via $X_{ij} =$

$\exp(w_{ij})/\sum_{k=1}^{p}\exp(w_{ik})$. Because a few taxa tend to dominate the community, while the rest are rare, we let $\theta_j = \log(p/2)$ for $j = 1, \ldots, 5$ and $\theta_j = 0$ otherwise. We consider $\Sigma = \mathbf{I}_p$, autoregressive of order 1 ($\rho = 0.2, 0.5$), or compound symmetric within groups ($\rho = 0.2, -0.2$) and autoregressive otherwise ($\rho = 0.2$). Outcomes are simulated from the linear log-contrast model (Equation 4.2) with errors $\epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$, choosing σ_ϵ^2 so that $R^2 = \text{var}(y_{true})/(\text{var}(y_{true}) + \sigma_\epsilon^2) \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$. We consider $(n, p) = (100, 100)$ and $(100, 200)$.

Four settings for β are considered, each assuming groups of size 5. In Setting A, we define β =(1, 0.6, 0.4, 0.5, 0.7, -0.6, -0.7, -0.4, -0.6, -0.9, 0, ..., 0). Each element in the first two groups is associated with the outcome, and the direction of effect is consistent within a group. In Setting B, we define β =(0.5, -0.9, 0.6, -0.7, 0.4, -0.5, 0.9, -0.6, 0.7, -0.4, 0, ..., 0). Each element in the first two groups is associated with the outcome, and the direction of effect varies within a group. In Setting C, we define β =(0.6, 0, 0.3, 0, 0.5, 0.3, -0.6, 0, -0.3, 0, -0.5, -0.3, 0, ..., 0) so that 2-3 of the five elements in each of the first three groups are nonzero. This mimics the case in which associated taxa do not strictly follow the predefined group structure. In Setting D, associated coefficients do not follow the grouping structure at all, with one element in each of the first four groups associated with the outcome (equal to -0.6, 0.6, -0.5, and 0.5, respectively).

In the phylogenetic scenario, our simulation strategy resembles that of [147]. We simulate datasets with $n = 100$ individuals. The OTU counts for each individual are generated from a Dirichlet-multinomial distribution, commonly used for both modeling and simulation of microbiome data because it accounts for overdispersion. The proportions and dispersion parameters are estimated from Charlson et al.'s upper-respiratory-tract microbiome dataset [19], which includes counts for 856 OTUs measured on 60 individuals. We generate 1000 OTU counts for each of the 100 simulated individuals using these estimated parameters. For computational convenience, we use the partitioning-around-medoids algorithm to aggregate the original 856 OTUs into $p = 120$ features in $q = 20$ groups of size ranging from 2-11 features. A group of size 7 (abundance 4.0%) and a group of size 5 (abundance 5.1%) are

associated with the outcome through the linear log-contrast model. Outcomes are again simulated from the linear log-contrast model (Equation 4.2) with errors $\epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$, choosing σ_ϵ^2 so that $R^2 \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$. Three settings are considered for the true coefficient vector. In Setting AA, every feature in the two associated groups has nonzero coefficients with consistent direction, with β =(0.6, 0.4, 0.7, 0.2, 0.3, 0.5, 0.3,-0.8, -0.5, -0.6, -0.4, -0.7, 0, ..., 0). In Setting BB, every feature in the two associated groups has nonzero coefficients with varying direction, with β =(0.6, -0.5, 0.7, 0.2, 0.3, -0.7, -0.4, 0.4, -0.8, 0.3, 0.5, -0.6, 0, ..., 0). In Setting CC, partial groups are associated, with β =(0.6, 0.4, 0, 0, 0.3, 0.5, 0, -0.8, 0, -0.6, -0.4, 0, ..., 0).

The methods to which we will compare CSGL are the lasso [121], group lasso [140], sparse group lasso with $\theta = 0.05$ and 0.95 [114], ridge regression [49], kernel-penalized regression [105], and the compositional lasso [74]. The unit-sum constraint is ignored for each of the non-compositional methods. For CSGL, we consider $\theta = 0$ (the "compositional group lasso", or CGL), 0.05 (CSGL05), and 0.95 (CSGL95), where higher θ corresponds to higher weight on the ℓ_1 penalty. In each case, 200 simulated datasets are generated; $\mu = 1$; and the regularization parameter λ is selected by the generalized information criterion as defined in Section 4.2.3.

4.4.2 Simulation Results

To evaluate the performance of the proposed method, we assess mean squared prediction error $||y - Z\hat{\beta}||_2^2/n$ on an independent data set of size n for each method (PE). In addition, for selected scenarios, we assess maximum estimation error, $\ell_\infty = \|\hat{\beta} - \beta\|_\infty$; model size; falsely selected coefficients, i.e., coefficients such that $\beta_j = 0$ but $\hat{\beta}_j \neq 0$ (FP); and falsely omitted coefficients, i.e., coefficients such that $\beta_j \neq 0$ but $\hat{\beta}_j = 0$ (FN).

Tables 4.1-4.3 present prediction, estimation, and selection accuracy results when the correlation structure is AR(1) with $\rho = 0.2$ and $\sigma = 0.5$, which corresponds to R^2 around 0.8 (although this varies by setting and is sometimes as high as 0.92). The extended prediction error results for this scenario are included in Appendix C Figure C.2.

Settings A and B have very similar results, so we present results for Settings B-D in the Tables. In Setting B (Table 4.1), when entire groups are associated with the outcome, CGL and CSGL05 have substantially lower prediction and estimation error than CL, which in turn has lower error than the non-compositional methods. CSGL95 behaves similarly to CL, although still with somewhat lower prediction error. Maximum estimation error follows the same patterns. These differences are more pronounced when $p = 200$ than when $p = 100$. The number of false negative coefficients is zero or negligible for all methods when $n = 100$ and $p = 100$; when $p = 200$, the number of false negative coefficients increases for the lasso, sparse group lasso ($\theta = 0.95$), and to a lesser extent the compositional lasso and CSGL ($\theta = 0.95$). No false negative coefficients were seen in CGL or CSGL05 with either $p = 100$ or $p = 200$. The superior performance of CGL and CSGL05 in Setting B occurs because $\theta = 0$ and $\theta = 0.05$ place (almost) all of the weight on the ℓ_2 penalty at the expense of the ℓ_1 penalty, and since the truly nonzero coefficients perfectly conform to the predefined group structure, the ℓ_2 penalty alone would be sufficient to select the correct features.

In Setting C, approximately half of the features in each of three groups are associated with the outcome, so some selection within associated groups is necessary to recover the true set of associated features. For CSGL, that corresponds to θ moderately close to 1, i.e., CSGL95 ($\theta = 0.95$). For $n = 100$ and $p = 100$, CGL, CSGL05, and CSGL95 all have similar prediction and estimation error, but the methods with larger θ choose a smaller model, which is consistent with the increasing weight on the ℓ_1 penalty inducing within-group sparsity. CGL often chooses a model of size 15, with seven false positive coefficients. Since three groups of size 5 are partially associated (8 of the 15 features are associated) and CGL must include or exclude whole groups, this is the closest CGL can get to the truth without omitting truly nonzero coefficients. When $p = 200$, CSGL95 has the lowest prediction error, estimation error, model size, and number of false positive and false negative coefficients, although CL performs similarly to CSGL95 by each of these metrics in this setting.

When associated taxa follow no grouping structure (Setting D), the prediction error and estimation error are similar for CL and CSGL95, which weights the ℓ_1 penalty more highly

than the ℓ_2 penalty (Table 4.3). The error for other versions of CSGL is substantially higher than that of CL because either the model encourages inclusion of extra coefficients in each group or, more often, misses the non-zero coefficient in each group due to the preponderence of zero coefficients.

Figures C.1-C.6 display prediction error across several correlation structure scenarios and signal to noise ratios, additionally including KPR and ridge regression for comparison. Across all correlation scenarios considered, the pattern of sparsity matters much more than compositionality or phylogenetic structure. That is, in settings where grouped coefficients are nonzero, methods with more weight on the L2 penalty have the lowest PE (CSGL95, SGL95), followed by methods weighting an L1 penalty more heavily (CSGL05, SGL05), followed by methods that do not induce sparsity (KPR, ridge), often with very little difference between the compositional and non-compositional varieties of these methods. Similarly, in settings where partially-grouped coefficients are nonzero, methods with more weight on the L1 penalty have lowest PE, followed by methods weighting an L2 penalty more heavily, followed by KPR and ridge regression. KPR demonstrates a substantial advantage over ridge regression when data are phylogenetically simulated and the kernel for KPR is based on patristic distances, although when the outcomes are simulated from the linear log-contrast model with sparse true coefficient vectors, the lasso-type methods still tend to have lower PE. Use of Aitchison's total variation matrix for KPR does not seem to improve prediction over ordinary ridge regression in the non-compositional scenarios considered here. Importantly, the model used for simulation was a sparse linear log-contrast model; results could differ depending on the assumed structure of the association.

4.5 Application to Gut Microbiota and BMI

We apply the proposed method to 16S sequencing data for the gut microbiome from the American Gut Project (www.americangut.org). Operational taxonomic units (OTUs) were clustered at 97% similarity and aligned to the GreenGenes reference taxonomy, and blooming bacteria filters were applied [5]. We included samples from subjects at least 18 years old with

Table 4.1: Simulation results for Setting B, in which full groups are associated with the outcome with differing direction of effect. Values in the table are Mean (SD) across simulated datasets.

(n, p)	Method	PE	ℓ_∞	Model Size	FP	FN
(100, 100)	L	0.58 (0.24)	0.31 (0.08)	19.2 (4.3)	9.3 (4.2)	0.1 (0.5)
	GL	0.53 (0.21)	0.29 (0.09)	11.9 (3.3)	1.9 (3.3)	0.0 (0.0)
	SGL05	0.53 (0.16)	0.30 (0.08)	12.1 (3.6)	2.1 (3.6)	0.0 (0.0)
	SGL95	0.62 (0.39)	0.32 (0.11)	17.8 (4.8)	8.0 (4.4)	0.2 (1.0)
	CL	0.51 (0.17)	0.28 (0.07)	17.4 (3.9)	7.4 (3.9)	0.0 (0.2)
	CGL	0.45 (0.17)	0.25 (0.07)	11.1 (2.5)	1.1 (2.5)	0.0 (0.0)
	CSGL05	0.45 (0.16)	0.25 (0.07)	11.2 (2.7)	1.2 (2.7)	0.0 (0.0)
	CSGL95	0.51 (0.16)	0.28 (0.07)	16.1 (3.5)	6.1 (3.5)	0.0 (0.1)
(100, 200)	L	1.46 (0.95)	0.55 (0.20)	11.8 (7.1)	4.8 (4.3)	3.0 (3.4)
	GL	0.70 (0.52)	0.35 (0.13)	11.1 (3.1)	1.4 (2.7)	0.3 (1.5)
	SGL05	0.71 (0.51)	0.35 (0.14)	11.1 (2.8)	1.3 (2.4)	0.2 (1.3)
	SGL95	1.28 (0.87)	0.51 (0.19)	12.6 (6.4)	4.8 (3.9)	2.2 (3.1)
	CL	0.85 (0.59)	0.39 (0.15)	15.7 (5.3)	6.6 (3.8)	0.9 (2.2)
	CGL	0.48 (0.14)	0.26 (0.07)	10.8 (2.1)	0.8 (2.1)	0.0 (0.0)
	CSGL05	0.48 (0.13)	0.26 (0.07)	10.8 (2.0)	0.8 (2.0)	0.0 (0.0)
	CSGL95	0.80 (0.54)	0.38 (0.14)	14.9 (4.6)	5.6 (3.3)	0.8 (2.0)

Table 4.2: Simulation results for Setting C, in which partial groups are associated with the outcome. Values in the table are Mean (SD) across simulated datasets.

(n, p)	Method	PE	ℓ_∞	Model Size	FP	FN
(100, 100)	L	0.47 (0.14)	0.29 (0.08)	12.7 (3.7)	5.3 (3.3)	0.6 (0.9)
	GL	0.60 (0.25)	0.36 (0.12)	16.3 (3.7)	8.7 (3.3)	0.4 (1.1)
	SGL05	0.57 (0.24)	0.35 (0.12)	16.6 (3.8)	9.0 (3.5)	0.4 (1.1)
	SGL95	0.46 (0.13)	0.29 (0.08)	12.8 (3.5)	5.2 (3.2)	0.5 (0.8)
	CL	0.41 (0.11)	0.20 (0.06)	11.4 (2.7)	3.5 (2.6)	0.1 (0.4)
	CGL	0.43 (0.17)	0.23 (0.07)	15.6 (2.6)	7.7 (2.2)	0.1 (0.8)
	CSGL05	0.42 (0.08)	0.22 (0.06)	15.3 (2.1)	7.3 (2.1)	0.0 (0.0)
	CSGL95	0.4 (0.09)	0.20 (0.05)	11.4 (2.4)	3.4 (2.4)	0.0 (0.2)
(100, 200)	L	0.62 (0.21)	0.39 (0.11)	11.6 (4.4)	5.0 (3.6)	1.4 (1.4)
	GL	0.74 (0.33)	0.44 (0.13)	15.5 (4.5)	8.6 (3.9)	1.1 (1.6)
	SGL05	0.74 (0.34)	0.44 (0.13)	15.2 (3.9)	8.4 (3.2)	1.2 (1.6)
	SGL95	0.62 (0.22)	0.39 (0.12)	11.7 (4.7)	5.0 (3.9)	1.4 (1.4)
	CL	0.50 (0.16)	0.25 (0.07)	10.9 (2.9)	3.2 (2.5)	0.3 (0.7)
	CGL	0.65 (0.48)	0.30 (0.13)	13.1 (5.4)	6.2 (2.7)	1.1 (2.8)
	CSGL05	0.61 (0.44)	0.29 (0.12)	13.4 (4.9)	6.3 (2.6)	0.9 (2.5)
	CSGL95	0.48 (0.18)	0.24 (0.07)	10.9 (2.8)	3.1 (2.5)	0.2 (0.9)

no prior diagnosis of irritable bowel disease or diabetes, with recorded age, sex, and BMI, and with at least 1000 total reads. The OTUs were aggregated at the family level, excluding singleton families, and groups were defined at the class level. Zeros in the OTU count matrix were replaced with the minimum observed proportion divided by two. For our primary analysis, we used subjects with samples run on January 4, 2017. This data set included abundance of 133 bacterial families for 405 subjects. We used CSGL with $\theta = 0.95$, which allows more extensive within-group selection by placing higher weight on the ℓ_1 penalty, and adjusted for age and sex by using the residuals from a linear model regressing BMI on age and sex as the outcome in the CSGL model. λ was selected by 10-fold cross validation. 26 of the 133 families had nonzero coefficient estimates (Table 4.4).

The families and classes selected by CSGL are broadly consistent with previously posited associations. For example, Actinobacteria are generally enriched in obesity [125]. Within Actinobacteria, the family Coriobacteriaceae is considered a "pathobiont," a taxon that is

Table 4.3: Simulation results for setting D, in which ungrouped coefficients are associated with the outcome. Values in the table are Mean (SD) across simulated datasets.

(n, p)	Method	PE	ℓ_∞	Size	FP	FN
(100, 100)	L	0.35 (0.07)	0.21 (0.06)	6.6 (1.9)	2.6 (1.9)	0.0 (0.0)
	GL	0.93 (0.43)	0.47 (0.14)	14.9 (6.8)	12.2 (5.6)	1.3 (1.4)
	SGL05	0.88 (0.43)	0.45 (0.14)	15.2 (6.7)	12.3 (5.5)	1.2 (1.4)
	SGL95	0.36 (0.07)	0.22 (0.06)	6.8 (2.3)	2.8 (2.3)	0.0 (0.0)
	CL	0.34 (0.06)	0.19 (0.04)	5.7 (1.6)	1.7 (1.6)	0.0 (0.0)
	CGL	1.11 (0.53)	0.47 (0.17)	7.3 (9.6)	5.8 (7.7)	2.5 (1.9)
	CSGL05	1.03 (0.54)	0.45 (0.18)	8.3 (9.5)	6.5 (7.5)	2.3 (2.0)
	CSGL95	0.35 (0.06)	0.19 (0.04)	5.7 (1.8)	1.7 (1.8)	0.0 (0.0)
(100, 200)	L	0.39 (0.10)	0.24 (0.06)	6.8 (1.9)	2.8 (1.9)	0.0 (0.0)
	GL	1.18 (0.40)	0.55 (0.09)	11.2 (5.5)	9.2 (4.5)	2.0 (1.1)
	SGL05	1.13 (0.38)	0.54 (0.10)	11.6 (5.6)	9.4 (4.7)	1.9 (1.1)
	SGL95	0.40 (0.11)	0.25 (0.06)	7.2 (2.5)	3.2 (2.5)	0.0 (0.0)
	CL	0.36 (0.08)	0.21 (0.05)	5.7 (1.7)	1.7 (1.7)	0.0 (0.0)
	CGL	1.46 (0.29)	0.59 (0.07)	0.9 (3.6)	0.7 (2.9)	3.8 (0.7)
	CSGL05	1.42 (0.33)	0.58 (0.09)	1.4 (4.7)	1.1 (3.7)	3.7 (1.0)
	CSGL95	0.37 (0.09)	0.21 (0.05)	5.6 (1.6)	1.6 (1.6)	0.0 (0.0)

innocuous or even beneficial under normal circumstances, but under particular environmental conditions, may contribute to a variety of diseases. Members of this family are important in host lipid metabolism [27], have been associated with metabolic dysfunction [69, 146], and are reduced with weight loss [86]. Among the Firmicutes, higher abundance of Christensenellaceae is associated with lower body mass index [41, 45], and addition of this bacterial family reduced weight gain in a mouse model [40]. Among the Proteobacteria, Burkholderiales and Pasteurellaceae were found in higher abundance in the salivary microbiome of lean individuals than obese [137], and Pasteurellaceae is associated with lower triglyceride levels [41]. Interestingly, Flavobacteriia, which here is found to be positively associated with BMI, also increases post-gastric bypass [52] and is found in higher abundance in saliva samples of lean individuals [137].

One widely-discussed association between the gut microbiome and obesity is the Firmicutes to Bacteroidetes ratio. In a number of mouse and human studies, a high Firmi-

cutes/Bacteroidetes (F:B) ratio was associated with obesity [71, 126, 145]. However, findings with respect to that ratio have varied among previous studies [86, 123], with some studies finding no effect or even the opposite association [36, 110]. In our CSGL model, the only nonzero Bacteroidetes coefficient is negative, which is consistent with a reduction of Bacteroidetes in higher-weight individuals. However, the Firmicutes coefficients are mixed, with generally positive coefficients for Bacilli and generally negative associations with Clostridia. A similar observation of discordant Firmicutes effects, and in particular an association between Clostridia and weight loss or lower weight, has been observed in previous studies cited as contradictory to the typical pattern of high F:B ratio [92, 106]. Armougom *et al.* find decreased Bacteroidetes and decreased Firmicutes at the phylum level, but increased Lactobacillus species (family Lactobacillaceae, phylum Firmicutes) with obesity [7]. Our results therefore support the growing evidence that diversity of effect within Bacteroidetes and Firmicutes, along with different dietary patterns and study populations, may explain the inter-study variation regarding the association of the F:B ratio with obesity.

4.6 Discussion

We have proposed a compositional sparse group lasso, which uses a linear log-contrast model to account for the compositionality of microbiome data, incorporates extrinsic phylogenetic or functional grouping structure and groupwise sparsity through an ℓ_2 penalty, and encourages sparsity within groups with an ℓ_1 penalty. By fitting the model at the feature level and accounting for group structure through a penalty on feature-level coefficients, CSGL leverages the similarity in the association of related taxa with an outcome, but has the flexibility to accommodate cases in which closely related taxa have somewhat different effects.

Tree-guided Automatic Subcomposition Selection Operator (TASSO) is an alternative approach that applies a tree-structured fusion penalty to the compositional lasso to incorporate phylogenetic structure. TASSO aims to select associated branches of the phylogenetic tree, choosing which level of taxon to include via variable fusion [122, 131]. CSGL differs from TASSO because we apply an ℓ_2 norm to groups of lower-level taxa, rather than effec-

Table 4.4: Bacterial families associated with BMI in the American Gut Project 16S fecal samples.

Phylum	Class	Order	Family	Coefficient
[Thermi]	Deinococci	Deinococcales	Deinococcaceae	-0.128
[Thermi]	Deinococci	Thermales	Thermaceae	0.662
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	-0.009
Actinobacteria	Actinobacteria	Actinomycetales	Nocardiaceae	0.056
Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	0.149
Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	0.127
Bacteroidetes	Bacteroidia	Bacteroidales	S24-7	-0.074
Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	0.299
Cyanobacteria	4C0d-2	YS2	[Unknown]	-0.192
Firmicutes	Bacilli	Gemellales	Gemellaceae	-0.055
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	0.001
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	0.168
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	0.268
Firmicutes	Bacilli	Lactobacillales	[Unknown]	0.096
Firmicutes	Clostridia	Clostridiales	Christensenellaceae	-0.077
Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	-0.116
Firmicutes	Clostridia	Clostridiales	Peptococcaceae	-0.034
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	0.001
Firmicutes	Clostridia	SHA-98	[Unknown]	-0.025
Firmicutes	Clostridia	Thermoanaerobacterales	Caldicellulosiruptoraceae	-0.072
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	-0.495
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	-0.081
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	-0.164
Tenericutes	Mollicutes	RF39	[Unknown]	-0.064
Tenericutes	RF3	ML615J-28	[Unknown]	-0.050
TM7	TM7-3	[Unknown]	[Unknown]	-0.168

tively applying an ℓ_1-norm to the corresponding aggregate taxon at a higher level on the tree. Another approach is the zero-sum elastic net [4], which only gives sparsity at the feature level. In contrast to both of these methods, CSGL allows estimation at the feature level with selection at both feature and group level.

The selection accuracy, estimation error, and prediction error of CSGL are consistently slightly superior to corresponding non-compositional methods. In the majority of simulation settings, CSGL also outperforms the compositional lasso by both measures, with the most noticeable differences when the grouping structure is most accurate with respect to truly associated coefficients. The compositional lasso may be preferred when there is no relevant grouping in the associated features, but choosing $\theta = 0.95$ for CSGL gives nearly equivalent performance in such situations. The pattern of sparsity relative to assumed group structure is more important than the compositionality constraint across all simulation scenarios.

As currently formulated, CSGL is restricted to including information at two levels (group and feature). When phylogenetic grouping is used, it is unclear which levels should be included. Analysis at the genus and class level, for example, seems just as reasonable as family and phylum level analysis. Sometimes scientific hypotheses or literature consensus provide clear direction. When this is not the case, or when taxonomic level of association is part of the scientific question, prediction error on a test set may be used to compare a small number of choices of levels. When functional grouping is of interest, the non-overlapping group structure of CSGL requires a single major function to be assigned to each taxon. Future work is needed to allow overlapping groups so that all of a taxon's functional components may be accounted for in the grouping structure.

Chapter 5

MULTILEVEL COMPOSITIONAL LASSO

In the previous chapter, we proposed a sparse group penalty for the linear log-contrast model. Since evolutionarily or functionally similar groups of taxa are often related similarly to phenotypes, incorporation of group information in penalized regression models improves estimation and prediction. However, the use of these models for prediction in new data implicitly assumes that the feature set is the same for future observations or datasets as in the training data. This assumption fails for microbiome data. In this chapter, we focus on the capacity to make predictions in new data based on the microbiome.

We propose a multilevel linear log-contrast model that includes both group-level and within-group relative abundances of taxa. Using this model in a new dataset, for taxa that were not previously observed or are less precisely identified (for example, identified at family rather than genus level), partial information may still be utilized via the group-level terms. ℓ_1 penalties on the group- and feature-level coefficients induces sparsity at both levels. The penalized form can be formulated as a convex optimization problem and solved using a coordinate descent method of multipliers algorithm. We assess the performance of the model in simulation studies and several studies of the gut microbiome and body mass index.

5.1 Prediction with Differing Feature Sets

Clinical interest in microbiome studies often lies primarily in predictive analyses. Although the high dimensionality and inter-subject variability of the microbiota make prediction challenging [59], several studies have demonstrated that microbial profiles may be used to good effect as a diagnostic and prognostic biomarker. For example, abnormal vaginal flora in the first trimester of pregnancy is predictive of preterm birth [35], gut microbiome composition

can be used to diagnose advanced fibrosis in nonalcoholic fatty liver disease with high accuracy [78], and pairing assessment of the fecal microbiota with the standard fecal occult blood test can improve the sensitivity of colorectal cancer screening by 45% while maintaining specificity [141]. Therefore, improving methods for microbiome-based prediction is of great clinical relevance.

Several regularized regression and feature selection methods for the microbiome have recently been developed. These methods generally use a linear log-contrast model to account for the compositional nature of microbiome data and add penalties, sometimes informed by the phylogenetic structure of the microbiome, to induce sparsity or smoothness and improve prediction accuracy [4, 74, 105, 119, 130, 131, 138, 142]. However, a key assumption of all of these existing approaches is that the set of features is the same in the prediction set as in the data on which the model is built. While this is a reasonable assumption in most scientific settings — the same variables are included on new surveys as previous ones, or the same genetic variants are assessed — it does not generally hold for microbiome data, for two reasons. First, there may actually be different taxa present in different datasets. Due to the high inter-individual variability, rare taxa in particular may not appear in the model-building dataset but may be present in future individuals. In addition, there may be differences in the level of taxonomic assignment possible between datasets. For example, the same biological taxon may be identified at the genus level in one dataset but only the family level in another. These differences, which we will refer to as differential resolution, may be due to choice of sequencing region, differences in sequencing accuracy, or choice of primers [17, 65, 77, 128].

At its essence, this is a problem of missing information: either the full taxon assignment is missing, such as when the genus of a taxon is unknown in one dataset, or the estimated effect of a taxon is missing, such as when the taxon is only present in the prediction set. However, contrary to usual missing data settings, we have access to partial information in both of these cases due to the phylogenetic relationships within the microbiome. Because phylogenetically similar taxa often have similar functional capabilities and environmental

preferences, it is reasonable to expect some shared effect among phylogenetically grouped taxa [9, 22, 51, 79, 89, 98, 131]. To leverage this partial information, we propose the multilevel compositional lasso, which fits a linear log-contrast regression model at two taxonomic levels and performs feature selection at each level. The group-level terms allow partial data such as class-level effects to be incorporated in the model; the within-group terms specialize this effect to particular genera or species, but can be omitted when the higher-resolution data are not available for particular taxa. Therefore, this model accounts for the compositionality of the data and allows more of the available information to be used for model-building and prediction. The multilevel compositional lasso is a special case of the constrained lasso [53], and similar algorithms and optimality results apply [42, 74, 113].

The remainder of the chapter is structured as follows. In Section 5.2, we describe the proposed multilevel compositional lasso (MCL) in detail, express it as a constrained convex optimization problem, and describe the coordinate descent method of multipliers algorithm used to solve it. We also comment on situations in which theoretical optimality properties will hold. Simulation results demonstrating the efficacy of the method are presented in Section 5.3. In Section 5.4, we apply the method to explore the association between the gut microbiome and BMI, and Section 5.5 concludes with a discussion.

5.2 Methods

5.2.1 Linear Log-Contrast Model

Although the linear log-contrast model was introduced in the previous chapter, we review it here to set the stage for the multilevel analogue. Suppose that taxon counts $\check{X}_{ij}$ are observed for taxa $j = 1, \ldots, p$ and subjects $i = 1, \ldots n$. Due to differences in total read count across samples, each individual's set of taxon counts is normalized to taxon proportions via $\widetilde{X}_{ij} = \check{X}_{ij} / \sum_{j=1}^{p} \check{X}_{ij}$. Following this normalization step, the data are compositional: observations for each subject are constrained to sum to one, so the rows of $\tilde{X}$ lie in a $(p-1)$-dimensional simplex. Compositionality induces negative correlation among features,

rendering traditional regression methods inappropriate.

To address this problem in the context of mixture data, Aitchison and Bacon-Shone proposed the linear log-contrast model [3]. In recent years this model has been widely applied to microbiome compositions [66, 74, 81, 113]. The linear log-contrast model relies on the log-ratio transformation [2], in which a reference component is selected and the log-ratios of each non-reference component to that reference are used as the predictors in a linear model. Suppose without loss of generality that component p is selected as the reference. Then the model is

$$y = X^p \beta_{\backslash p} + \epsilon$$

where X^p is the $n \times (p-1)$ matrix with elements $X^p_{ij} = \log(\widetilde{X}_{ij}/\widetilde{X}_{ip})$, $\beta_{\backslash p}$ is a $(p-1)$-dimensional vector of coefficients, and $\epsilon \sim N(0, \sigma^2)$. By defining $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, the model may be equivalently expressed in symmetric form via

$$y = X\beta + \epsilon, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0$$

where $X = \log(\widetilde{X})$ are log-proportions. Due to the high dimensionality of microbiome data, applying combinations of penalties to this model (possibly informed by phylogenetic structure) often results in superior estimation and prediction accuracy [74, 113, 130].

5.2.2 Multilevel Linear Log-Contrast Model

Aitchison's linear log-contrast model can accommodate only a single-level composition. When microbiome data are summarized at just one level of phylogenetic classification, such as genus or family level, this is appropriate. However, taxon abundance matrices are accompanied by a phylogenetic tree, and taxa may be identified at each level from kingdom to species. Within each higher-level group, there is a composition at the lower level, and it is often unclear which phylogenetic level is most relevant in a particular setting. One solution to this is to always model at the lowest sensible level (e.g., genus) and assume that higher-level effects, such as associations with an entire class or phylum, will be distributed across each

genus within that higher taxonomic category. However, if the association is at a higher level, the distributed effect size for individual genera may be small and difficult to detect. In contrast, always modeling at higher levels of taxonomy negates the ability to detect fine-scale associations.

The tree-structured fused lasso penalty accounts for the multiplicity of potential levels by shrinking differences between coefficients in a single phylogenetic branch to zero. It thereby adaptively chooses a single level of phylogeny within each group, allowing the selected level to differ between groups [130]. The biological assumption behind this method is that phylogenetically similar taxa have similar effects. This is generally the case [9, 89, 98, 79], but there are important exceptions. For example, a high Firmicutes to Bacteroidetes ratio is widely cited as being predictive of obesity [28, 71, 126, 145], but other studies find no association or even the opposite effect [86, 123]. This may be due to lower-level taxa within each phylum with differing strength and direction of association with obesity [92, 106]. The same idea holds at much finer phylogenetic resolutions: although they are in the same genus, *Bacillus anthracis* is pathogenic, while *Bacillus cereus* is not [77]. A method that penalizes coefficients to encourage similar effect estimates in closely-related taxa may not accommodate these situations. It is therefore important to allow differences in association within phylogenetic groups.

In addition, when primary scientific interest lies in prediction based upon new data (for example, building a microbiome-based biomarker for disease diagnosis or prognosis), microbiome data poses more fundamental analytic challenges. Choices made in study design have several important effects on composition. First, in a 16S rRNA sequencing protocol, often only two or three out of the nine hypervariable regions are sequenced, and different hypervariable regions permit differing accuracy and depth of taxonomic identification [77]. The choice of sequencing region may therefore strongly affect which taxa are found to be dominant [65]. Even for the same hypervariable region, the choice of a particular "universal" primer may result in detecting or overlooking entire genera of bacteria [128]. As a result, two studies examining the same body site in the same study population may identify different

taxa, and may do so at differing taxonomic levels of resolution, simply due to choice of primers and hypervariable regions. Alongside these technical differences, since rare taxa may not be observed in every dataset, there are also likely to be true biological differences between feature sets in different studies.

Prediction models that assume the same set of features in the model-building dataset as the prediction set are not immediately applicable when feature sets differ across studies due to biology or differential resolution. Application of standard models for prediction with differing feature sets requires either fitting the model at a high enough phylogenetic level that most or all taxa are shared between datasets, or fitting the model at a lower phylogenetic level, then excluding taxa in the prediction set that are not observed in the model-building set. Both approaches result in a loss of information, due in the first case to aggregating taxa with potentially different effects and in the second case to excluding taxa entirely.

To address the concerns of differing feature sets between studies while allowing potentially differing effects within phylogenetic groups, we propose a multilevel linear log-contrast model that incorporates both group-level effects and within-group modifications of that effect. Assume that extrinsic grouping information based on phylogenetic similarities among taxa is known, so that the p taxa are divided into q groups of size $p_1, ..., p_q$. The group-level proportions are based on aggregate counts within each group; within-group abundances modulate the group-level effect, accommodating the situation in which closely related taxa have differing associations with a host phenotype. For prediction in new data, if group membership is known but precise taxon classification is not, the within-group effect may be excluded, and the taxon still contributes information through the overall group effect.

In order to describe the model more precisely, we first introduce some notation. We will refer to elements of $\widetilde{X}$, the matrix of overall proportions, as $\widetilde{X}_{ijk}$. Here i indexes the rows of $\widetilde{X}$ (individuals); $j = 1, \ldots, q$ indexes blocks of columns (group membership); and $k = 1, \ldots, p_j$ indexes features (taxa) within each group. We then define the $n \times q$ matrix of group-level proportions $\widetilde{Z}$ with elements $\widetilde{Z}_{ij} = \sum_{k=1}^{p_j} \widetilde{X}_{ijk}$ and the $n \times p$ matrix of within-group proportions W with elements $\widetilde{W}_{ijk} = \frac{\widetilde{X}_{ijk}}{\sum_{k=1}^{p_j} \widetilde{X}_{ijk}} = \frac{\widetilde{X}_{ijk}}{\widetilde{Z}_{ij}}$. The corresponding matrices

of log-proportions are designated $Z = \log(\widetilde{Z})$ and $W = \log(\widetilde{W})$. Group-level coefficients are represented by $\beta = (\beta_1, \ldots, \beta_q)$ and within-group coefficients by $\gamma = (\gamma^{(1)}, \ldots, \gamma^{(q)})$, where $\gamma^{(j)}$ are the coefficients corresponding to group j. The proposed multilevel linear log-contrast model is

$$y = Z\beta + W\gamma + \epsilon \quad \text{s.t. } \sum_{j=1}^{q} \beta_j = 0 \text{ and } \sum_{k=1}^{p_j} \gamma_k^{(j)} = 0 \quad \forall j.$$

The zero-sum constraints on each coefficient vector or subvector make this symmetric form equivalent to modeling between- and within-group log-ratios of relative abundances, thereby eliminating the unit-sum constraint on the relative abundances. This model can use partial information for novel or lower-resolution taxa in the prediction set, based on the observation that phylogenetic groups of taxa often have similar effects, by including the taxon in its higher-level phylogenetic group. When lower-level information is available, within-group proportions for individual taxa provides the flexibility to model closely related taxa that have differing associations with the host phenotype.

5.2.3 Multilevel Compositional Lasso

Because microbiome data tends to be high-dimensional, with more taxa observed than subjects, the multilevel linear log-contrast model is usually not identifiable. Many choices for penalty functions exist that yield different patterns of smoothness and sparsity for coefficient estimates. We choose ℓ_1 penalties for both group-level and within-group coefficients, promoting sparse estimates for both group-level and within-group features. In addition to reducing prediction error and aiding in model identifiability, sparsity-inducing penalties identify which taxa are primarily responsible for the association with a phenotype, a question of great scientific importance.

Adding ℓ_1 penalties on within-group and between-group parameters to the loss function yields the multilevel compositional lasso (MCL), defined by

$$\ell(\beta, \gamma) = \frac{1}{2n}\|y - Z\beta - W\gamma\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\gamma\|_1 \tag{5.1}$$

$$\text{s.t. } \sum_{j=1}^{q} \beta_j = 0 \text{ and } \sum_{k=1}^{p_j} \gamma_k^{(j)} = 0 \quad \forall j.$$

When λ_2 is large relative to λ_1, $\hat{\gamma}$ will be highly sparse and group-level effects $\hat{\beta}$ will dominate the model. Conversely, λ_1 large relative to λ_2 will tend to encourage more nonzero $\hat{\gamma}$. We optimize over a grid of (λ_1, λ_2), using cross-validation to select the optimal lambda. This model can be expressed as a special case of the constrained lasso by stacking the data and constraint matrices. The constrained lasso model would be

$$\begin{aligned} \ell(\beta, \gamma) = \frac{1}{2n} \left\| y - \begin{bmatrix} Z & W \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \theta\beta \\ (1-\theta)\gamma \end{bmatrix} \right\|_1 \\ \text{s.t. } \sum_{j=1}^{q} \beta_j = 0 \text{ and } \sum_{k=1}^{p_j} \gamma_k^{(j)} = 0 \quad \forall j \end{aligned} \tag{5.2}$$

where the two penalty parameters λ_1 and λ_2 have been re-expressed $\lambda_1 = \theta\lambda$ and $\lambda_2 = (1-\theta)\lambda$.

An coordinate descent method of multipliers algorithm is used to solve problem (5.1). The constraints can be re-written in matrix form as $B\beta + C\gamma = 0$ where $B = [\mathbf{0}_{q\times q}, \mathbf{1}_q]^\top$ and $C = [D_{p\times q}, \mathbf{0}_p]^\top$. Here, D is a $p \times q$ matrix with column vectors $\mathbf{1}_{p_j}$ on the diagonal and 0 elsewhere. Using the matrix form for the constraints, we form the augmented Lagrangian,

$$\mathcal{L}_\mu(\beta, \gamma, \xi) = \frac{1}{2n}\|y - Z\beta - W\gamma\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\gamma\|_1 + \frac{\mu}{2}\|B\beta + C\gamma + \xi\|_2^2 - \frac{\mu}{2}\|\xi\|_2^2 \tag{5.3}$$

We then iteratively update β, γ, and ξ by coordinate descent as follows:

1. Initialize $\beta = \mathbf{0}_q$, $\gamma = \mathbf{0}_p$, $\xi = \mathbf{0}_{q+1}$, $\mu > 0$.

2. Until convergence, iterate:

   (a) For $j = 1, \ldots, q$ update β_j via

$$\beta_j \leftarrow \frac{1}{Z_j^\top Z_j + \mu} \cdot S\left(Z_j^\top (y - W\gamma - Z_{-j}\beta_{-j}) - \mu(\sum \beta_{-j} + \xi_{q+1}), \lambda_1 \right)$$

(b) For $i = 1, \ldots, p$ update $\gamma_i = \gamma_k^{(j)}$ via

$$\gamma_k^{(j)} \leftarrow \frac{1}{(W_j^{(k)})^\top W_j^{(k)} + \mu} \cdot S\left((W_k^{(j)})^\top r_{-jk} - \mu(\sum \gamma_{-k}^{(j)} + \xi_j), \lambda_2\right)$$

where r_{-jk} are the residuals without the ith term, $(y - Z\beta - W^{(-j)}\gamma^{(-j)} - W_{-k}^{(j)}\gamma_{-k}^{(j)})$.

(c) Update ξ via

$$\xi \leftarrow \xi + \mu \cdot \left[\sum_{k=1}^{p_1} \gamma_k^{(1)}, ..., \sum_{k=1}^{p_q} \gamma_k^{(q)}, \sum_{j=1}^{q} \beta_j\right]^\top$$

If any groups contain only one feature, we exclude that feature from W and γ, estimating the effect only at the group level. The within-group proportions of such a feature would be uniformly 1, leading to $W_{ik} = 0$ for $i = 1, ..., n$, and γ_k associated with a feature that is uniformly zero is not estimable.

Under some assumptions, optimality results and inference procedures for other constrained ℓ_1 penalized regression models hold for MCL. In the setting of the compositional lasso (CL) [74], which is equivalent to MCL with only the group-level term, with high probability there exists an optimal solution that satisfies sign consistency and bounded ℓ_∞ loss. The main condition for this result is an Irrepresentable Condition that requires low correlation between features with truly zero and nonzero coefficients. For these results to apply directly to MCL, within-group and group-level features must satisfy the same requirement of low correlation between zero and nonzero features.

To see when this would fail, consider the case of a single taxon changing drastically in abundance and driving the entire association with the host phenotype. Assume without loss of generality that neither this taxon nor its group is the reference taxon or group. We will refer to this hypothetical association-determining taxon as Species A in Family 1. For Species A, an increase in absolute abundance (with no change in abundance for the other taxa in its group) would result in an increase in relative abundance within Family 1, and correspondingly an increase in log-ratio of within-group proportion. Similarly, because all other taxa have constant abundance, the increase in abundance of Species A corresponds to

an overall increase in relative abundance of Family 1. Therefore, the features corresponding to Species A and Group 1 are correlated. By itself, this is still not problematic, since the Irrepresentable Condition only restricts correlation between zero and nonzero features. The condition is violated only under the additional assumption that the association between microbiota and outcome is due to Species A but *not* Family 1 or vice versa.

This perfect storm rarely occurs. In some conditions, a dramatic bloom of a single pathogen does drive the phenotype. For example, *Clostridium difficile* infection is a very serious condition driven purely by this bacterium. However, even in this case, it would not necessarily be inappropriate for the model to attribute some of the association of microbiota with disease to a phylogenetic rank somewhere above the species *C. difficile*, and a biologist or clinician looking at the data alongside the results from MCL could quite easily attribute any group-level effect to the correct species. Clinical observation and other types of models can also help identify settings in which a single, very specific taxon determines or strongly predicts a phenotype. More often, dysbiosis is characterized by smaller changes in many taxa, a setting in which correlation between the group-level and within-group features is minimal. Therefore, in most cases of scientific interest for feature selection and prediction with the microbiome, with high probability there exists an optimal solution for MCL that satisfies the same theoretical properties as the optimal solution for CL.

5.3 Simulation Study

We compare the performance of our method to CL and several non-compositional methods, simulating data from both the multilevel linear log-contrast model and the standard linear log-contrast model. Mean squared prediction error is computed both when the same set of features was observed in the validation set as the training set ("high-resolution") and when some features are not identified to the same level in the validation set as the training set ("low-resolution").

5.3.1 Simulation Methods

Taxon proportions are simulated as described in the previous chapter. Specifically, we first simulate $w_{n\times p} \sim N_p(\theta, \Sigma)$, then generate the matrix of taxon proportions $\widetilde{X}$ via $\widetilde{X}_{ij} = \exp(w_{ij}) / \sum_{k=1}^{p} \exp(w_{ik})$ so that $\widetilde{X}$ follows a logistic normal distribution [1]. We let $\theta_j = \log(p/2)$ for $j = 1, \ldots, 5$ and $\theta_j = 0$ otherwise, and Σ is autoregressive of order 1 with $\rho = 0.2$. Groups are all of size 5. Outcomes are simulated from either the linear log-contrast model or the multilevel linear log-contrast model, to assess the robustness of our method to different forms of association.

For outcomes generated from the linear log-contrast model using the overall taxon proportions $\widetilde{X}$, which we refer to as the compositional model, we consider two settings. In setting C1, nonzero coefficients partially correspond to group structure, with $\beta =$ (0.8, -0.4, -0.5, 0, 0, -0.8, 0.5, 0, 0, 0.4, 0, ..., 0). In setting C2, the associated coefficients are not grouped at all, with $\beta =$ (0.3, 0, 0, 0, 0, -0.4, 0, 0, 0, 0, 0.4, 0, 0, 0, 0, -0.3, 0, ..., 0). In both of these settings, the set of predictors used to generate the outcomes is different than the set of predictors included in the model, since the overall log-proportions X were used to generate outcomes via $Y = X\beta + \epsilon$ (with $\epsilon \sim N(0, \sigma^2)$), whereas group-level and within-group log-proportions were covariates in MCL.

For outcomes generated from the multilevel linear log-contrast model, which we refer to as the subcompositional model, we consider three settings. In setting SC1, there are both group-level and within-group effects, with $\beta =$ (-0.3, 0, 0.3, 0, ..., 0) and $\gamma =$ (-0.5, 0.8, -0.3, 0, 0, 0.5, -0.8, 0.3, 0, ..., 0). This is the most scientifically relevant situation, in which both overall group abundances and distributions of taxa within groups are associated with the host phenotype. Setting SC2 represents the case in which only lower-level taxonomic groups matter, with $\beta = 0_q$ and $\gamma =$ (-0.5, 0.8, -0.3, 0, 0, 0.5, -0.8, 0.3, 0, ..., 0). For example, this could represent a situation in which total relative abundance of Proteobacteria doesn't matter of its own accord, but having high Alphaproteobacteria relative to Gammaproteobacteria (within-group relative abundances) is predictive of a phenotype. Conversely, setting SC3

represents the case in which only upper-level groups matter, with coefficients β = (-0.5, 0, 0.5, 0, ..., 0) and $\gamma = 0_p$. Scientifically, this would correspond to a situation in which, for example, phylum-level relative abundances are associated with an outcome regardless of which mix of genera contribute to the total phylum abundance for each subject.

Comparison methods include the non-compositional lasso (L) [121], group lasso (GL) [140], and sparse group lasso (SGL) with $\alpha = 0.95$ [114], all fit using overall log-proportions and ignoring the constraint, as well as the compositional lasso (CL) [74]. In each simulation setting, the methods are compared at high resolution, when the set of features is the same in the prediction set as in the training set, and at low resolution, when the second, third, and fourth groups of features are considered to be observed only at group level. For each method except MCL, in the low-resolution case, outcomes are simulated from the original (high-resolution) model, then fit at lower resolution so that outcomes may be predicted from lower-resolution data. λ is selected based on prediction error in an independent dataset of size n.

5.3.2 Simulation Results

From the simulation results, we observe that in the high-resolution setting where data were simulated from the multilevel linear log-contrast model (the subcompositional model), prediction error is always lowest with MCL (Table 5.1). Generally, the group lasso is a close second, demonstrating the importance of incorporating structural information when relevant. When data were simulated from the ordinary linear log-contrast model (compositional model), prediction error is similar in MCL and the alternate methods when nonzero coefficients observe the pre-defined grouping structure at least partially, whereas when coefficients are ungrouped and data were simulated from this model, MCL has substantially higher prediction error. For ungrouped coefficients, the effect in MCL is divided between the group-level coefficient and the within-group coefficient, which in some cases causes both to be excessively shrunk towards zero and results in the increase in prediction error. This is similar to the theoretical situation discussed in the previous section.

In the low-resolution setting, MCL also always has prediction error similar to or lower than the alternatives (Table 5.2). Prediction error between MCL and all four alternative methods is fairly similar in setting C2, when there is no true group effect. In setting SC2, because only higher-level taxonomy matters, the methods that incorporate group structure (GL, SGL) have very similar prediction error to MCL, whereas CL has higher prediction error. That is, when the scientific aim is to build a model that may be used for prediction in independent microbiome datasets for which the set of observed features (taxa) differs, regardless of the true form of association, MCL provides the model with lowest prediction error in the new dataset.

Table 5.1: Simulation results for all settings, when the resolution in the prediction (validation) set is the same as that in the training set. Under "Model", SC indicates the subcompositional (multilevel) linear log-contrast model and C indicates the usual linear log-contrast model. Reported value is mean (SD) mean squared prediction error calculated on an independent dataset.

Model	(n, p)	Setting	L	GL	SGL	CL	MCL
SC	(100, 100)	SC1	0.38 (0.07)	0.35 (0.06)	0.37 (0.07)	0.37 (0.06)	0.34 (0.06)
		SC2	0.34 (0.06)	0.33 (0.06)	0.34 (0.06)	0.33 (0.06)	0.31 (0.05)
		SC3	0.32 (0.05)	0.29 (0.05)	0.31 (0.05)	0.36 (0.06)	0.27 (0.04)
	(100, 200)	SC1	0.41 (0.07)	0.37 (0.06)	0.40 (0.07)	0.41 (0.07)	0.35 (0.06)
		SC2	0.37 (0.07)	0.34 (0.06)	0.36 (0.06)	0.36 (0.06)	0.33 (0.05)
		SC3	0.32 (0.05)	0.30 (0.05)	0.32 (0.05)	0.37 (0.06)	0.27 (0.04)
C	(100, 100)	C1	0.35 (0.06)	0.33 (0.05)	0.34 (0.06)	0.32 (0.05)	0.33 (0.05)
		C2	0.31 (0.05)	0.37 (0.06)	0.31 (0.05)	0.30 (0.05)	0.42 (0.07)
	(100, 200)	C1	0.35 (0.06)	0.33 (0.06)	0.34 (0.06)	0.33 (0.06)	0.33 (0.06)
		C2	0.32 (0.06)	0.38 (0.07)	0.32 (0.06)	0.31 (0.06)	0.46 (0.08)

5.4 Gut Microbiome and BMI

We apply MCL to the association of BMI with the gut microbiome in one batch of American Gut Project (AGP) data and assess the results within that dataset. We then use the model to predict BMI in two other datasets: a second batch from the AGP, and the Human Microbiome Project (HMP) data. The second batch of AGP data represents a scenario in which there

Table 5.2: Simulation results for all settings, when the resolution in the prediction (validation) set is lower than that in the training set. Under "Model", SC indicates the subcompositional (multilevel) linear log-contrast model and C indicates the usual linear log-contrast model. Reported value is mean (SD) mean squared prediction error calculated on an independent dataset.

Model	(n, p)	Setting	L	GL	SGL	CL	MCL
SC	(100, 100)	SC1	1.17 (0.19)	1.15 (0.19)	1.17 (0.19)	1.17 (0.19)	1.05 (0.15)
		SC2	1.14 (0.18)	1.12 (0.18)	1.14 (0.18)	1.13 (0.19)	1.04 (0.14)
		SC3	0.28 (0.04)	0.27 (0.04)	0.27 (0.04)	0.31 (0.05)	0.27 (0.04)
	(100, 200)	SC1	1.17 (0.19)	1.14 (0.18)	1.16 (0.19)	1.17 (0.19)	1.05 (0.16)
		SC2	1.14 (0.19)	1.13 (0.18)	1.14 (0.18)	1.13 (0.18)	1.03 (0.16)
		SC3	0.27 (0.04)	0.27 (0.04)	0.27 (0.04)	0.31 (0.05)	0.27 (0.04)
C	(100, 100)	C1	1.25 (0.18)	1.21 (0.16)	1.24 (0.18)	1.24 (0.17)	1.14 (0.16)
		C2	0.67 (0.10)	0.67 (0.09)	0.67 (0.10)	0.67 (0.10)	0.65 (0.10)
	(100, 200)	C1	1.30 (0.22)	1.25 (0.21)	1.29 (0.22)	1.28 (0.22)	1.16 (0.18)
		C2	0.69 (0.10)	0.68 (0.10)	0.68 (0.10)	0.68 (0.10)	0.67 (0.10)

are few differences in feature sets due to technical or study design disparities, since the same investigators and analysis pipeline were used to produce the data. However, there are still differences in feature sets due to inter-subject variability. The HMP data represents the setting in which differential resolution due to technical or study design differences is expected in addition to biological variability. The biological differences in taxon sets is also expected to higher between AGP and HMP, since the HMP subjects are adults with irritable bowel diseases rather than the healthy adults included in the AGP analyses.

5.4.1 American Gut Project Analysis

We apply MCL to demultiplexed 16S sequencing data for the gut microbiome from the American Gut Project (http://qiita.microbio.me, FASTQ ID 48740, OTU matrix ID AG57, generated 03/11/2018). OTUs were clustered at 97% similarity and aligned to the GreenGenes reference taxonomy. We included samples from subjects at least 18 years old with no prior diagnosis of irritable bowel disease or diabetes, with available age, sex, and BMI, and with at least 1000 total reads. The OTU data was aggregated at the genus level, excluding

genera with fewer than 10 reads. Taxonomic groups were defined at the class level. Zeros in the OTU count matrix were replaced with 0.5, the maximum rounding error. The final dataset, which we will refer to AG57, included abundance of 274 bacterial genera for 424 subjects.

Of the 29 classes, 9 contained only one observed genus, and maximum group size was 63 genera. The nine singleton genera were excluded from the within-group feature matrix, W, and analyzed only at the group level. Therefore, we applied MCL with input data $Z \in \mathbb{R}^{(424\times 29)}$ and $W \in \mathbb{R}^{(424\times 265)}$. We set $\mu = 1$ and selected (λ_1, λ_2) by 10-fold cross validation. For comparison, we also fit the compositional lasso on the original log-proportions $X \in \mathbb{R}^{(424\times 274)}$ (CL-genus) and, separately, on the class-level log-proportions (CL-class), again selecting λ by 10-fold cross validation. We adjusted for age and sex by performing regression on age and sex in a first step, then using the residuals as the outcome measure for MCL and CL.

In some settings, it is possible to have two (λ_1, λ_2) pairs with nearly identical prediction error, one that emphasizes the contribution of within-group features and one that emphasizes group-level features. When that is the case, either a finer grid of λ values may be fit near the two optimal regions or scientific interest in group versus feature level effects may drive a preference for one pair over the other. However, in this case, the cross-validated prediction errors have a single minimum (Figure 5.4.1). The resulting model includes both group- and within-group effects (Tables 5.3 and 5.4).

Overall, the results are consistent with those found in the previous chapter, but comparing the class-level and genus-level coefficients provides additional insight into which level of taxonomy matters for the association between a particular taxon or group of taxa and BMI. For instance, Actinobacteria has the largest nonzero class-level association with BMI, with only two non-zero genus-level adjustments. The phylum and class of Actinobacteria have been consistently positively associated with obesity in previous studies, and this model suggests that these higher-level associations are appropriate for Actinobacteria [46, 125, 135, 145]. Also consistent with our findings, the Methanobacteria have been uniformly positively asso-

ciated with obesity [26, 43]. We estimate that Betaproteobacteria are negatively associated with BMI and Deltaproteobacteria are positively associated. While the Deltaproteobacteria association has been previously identified [26, 86], Betaproteobacteria are not often specifically discussed, although they seem to be associated with high sugar consumption in mice [111] and were found at very high abundance in malnourished children in Bangladesh [88]. The role of Betaproteobacteria may be worth investigating further in Western populations. The classes selected are very similar between MCL and CL-class.

In contrast, Bacilli and Clostridia each have no class-level effect, but each has several nonzero genus-level coefficients. The estimated direction of association with BMI varies within each group, suggesting that a class-level estimate would not be sufficient to capture the association. For example, higher abundance of genus rc4-4 in family Peptococcaceae relative to other Clostridia is associated with higher BMI, whereas lower abundance of genera in the family Mogibacteriaceae and order Clostridiales relative to other Clostridia is associated with higher BMI. Often, the overall association of obesity with Clostridia concentration is negative [43], but positive associations have been found with Lachnospiraceae and Peptococcaceae, as we saw [73, 85]. For the Bacilli, higher Bacillaceae is associated with higher BMI, and in contrast to the usual associations, higher Enterococcus is associated with lower BMI. However, this coefficient is for Enterococcus relative to other Bacilli, not the overall relative abundance of Enterococcus. Further investigation of the association of Bacilli subcompositions with obesity may clarify the role of particular genera.

5.4.2 Prediction in New Data

We predict BMI in two new datasets based on the MCL and CL models. Age and sex were adjusted for by including the previously fitted linear regression of BMI on age and sex as part of the MCL or CL predictions. In each new dataset, predictions based on the CL-class model were computed by aggregating the OTU table by class, excluding classes that were not observed in the original AG57 dataset, and using the remaining class-level relative abundances in the CL-class model from the AG57 data. Predictions based on the CL-genus

Table 5.3: Class-level coefficients from MCL and class-level CL. The number of non-zero genera per class refers to the MCL within-group coefficient estimates.

Phylum	Class	# Genera (# Nonzero)	MCL $\hat{\beta}$	CL $\hat{\beta}$ (Class-Level)
Actinobacteria	Actinobacteria	28 (2)	0.251	0.145
Actinobacteria	Coriobacteriia	6 (0)	·	0.007
Bacteroidetes	Bacteroidia	18 (0)	·	·
Bacteroidetes	Cytophagia	1	·	·
Bacteroidetes	Flavobacteria	3 (0)	·	·
Bacteroidetes	[Saprospirae]	1	·	·
Bacteroidetes	Sphingobacteriia	2 (0)	·	·
Cyanobacteria	4C0d-2	1	-0.181	-0.133
Cyanobacteria	Chloroplast	1	-0.214	-0.063
Deinococcus-Thermus	Deinococci	3 (0)	·	·
Elusimicrobia	Elusimicrobia	2 (0)	·	·
Firmicutes	Bacilli	36 (3)	·	·
Firmicutes	Clostridia	63 (9)	·	·
Firmicutes	Erysipelotrichi	12 (0)	·	·
Fusobacteria	Fusobacteriia	2 (0)	·	·
Lentisphaerae	[Lentisphaeria]	2 (0)	-0.153	-0.094
Proteobacteria	Alphaproteobacteria	21 (0)	·	·
Proteobacteria	Betaproteobacteria	20 (0)	-0.122	-0.017
Proteobacteria	Deltaproteobacteria	5 (0)	0.346	0.132
Proteobacteria	Epsilonproteobacteria	1	·	·
Proteobacteria	Gammaproteobacteria	31 (0)	·	·
Spirochaetes	[Brachyspirae]	1	·	·
Synergistetes	Synergistia	5 (0)	·	·
Tenericutes	Mollicutes	2 (0)	-0.015	-0.037
Tenericutes	RF3	1	-0.019	·
Verrucomicrobia	Opitutae	1	·	·
Verrucomicrobia	Verrucomicrobiae	1	·	·
Euryarchaeota	Methanobacteria	2 (0)	0.106	0.061
Euryarchaeota	Thermoplasmata	2 (0)	·	·

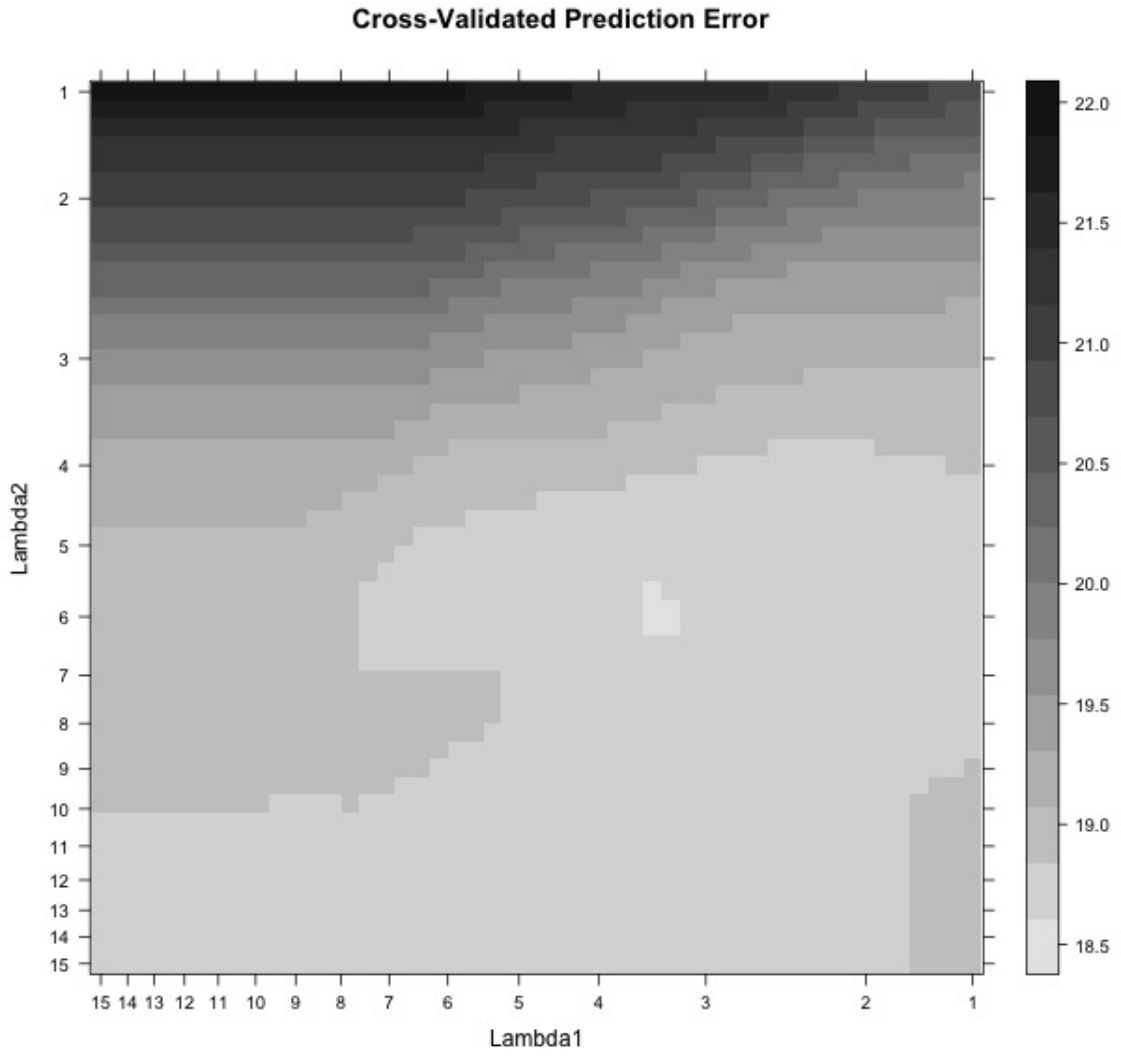


Figure 5.1: Cross-validated PE on a grid of λ_1, λ_2 values. Darker colors indicate higher mean squared prediction error. Large λ_2 values induce high sparsity in within-group effects γ, and large λ_1 values induce high sparsity in group-level effects β.

model followed the same pattern, but aggregating OTU abundances at the genus level. For predictions based on MCL, OTU abundances were aggregated at the genus level. Group assignment was made at the class level based on classes observed in the AG57 data, and the class-level aggregate proportions were generated as in CL-genus, including even the OTUs

for which genus-level information was unavailable. Genus-level proportions were computed within each class using all genera from that class. Then, genera that were not observed in the AG57 data were excluded, and the remaining genus-level relative abundances were included in the predictions from the MCL model. Hence all OTUs that belonged to classes observed in AG57, whether the genus was taxonomically identified or not, were included in the class-level term of the model. OTUs for which the genus was identified and previously observed were additionally included in the genus-level term of the model.

American Gut Project Validation

Our first validation set is another cohort from the American Gut Project data (FASTQ ID 45404, OTU matrix 45578, generated 02/09/2018). The data were processed in exactly the same way as AG57 except that we excluded singleton genera rather than genera observed fewer than 10 times. Of the 416 genera observed at least twice in the 285 AG59 subjects, 223 matched those in AG57, and 28 of the 29 AG57 classes were observed. The mean squared prediction error was highest using CL-genus, at 35.47. The prediction error using MCL was slightly lower than that of CL-class, at 34.36 and 34.45, respectively. In this dataset, fitting the model at the class level seems to be nearly as effective as including genus-level information, and the small number of nonzero within-group (γ) coefficients in MCL supports that conclusion. However, fitting the compositional lasso at genus level and attempting to apply in a dataset with different observed genera results in higher prediction error.

Human Microbiome Project Validation

To assess the performance of the MCL model in a dataset that less closely matches the model-building data, we also consider a subset of 30 adult subjects (age range: 19-74 years) from the Human Microbiome Project (HMP). Of these subjects, 13 have Crohn's disease, 9 have Ulcerative Colitis, and 8 do not have an irritable bowel disease. We again excluded singleton genera. Of the 348 observed genera, 140 were included in the original AG57 model. 22 of the 29 classes were observed in the HMP set. The prediction error was 30.8 for MCL,

42.4 for CL-genus, and 30.5 for CL-class. The difference between MCL prediction error and CL-genus is higher in this validation set, as we would expect in a dataset with more biological and technical differences in taxon detection and assignment. In this case, CL-class has slightly lower prediction error than MCL, suggesting that when there is relatively little overlap between feature sets and stronger demographic differences between study populations (here, based on health status, particularly the presence of irritable bowel diseases), fitting a predictive model at higher phylogenetic settings is likely to be sufficient.

5.5 Discussion

By fitting a regression model using both group-level compositions and within-group compositions, the multilevel linear log-contrast model provides flexibility in the face of unknown level of association, potentially differing direction and strength of association within upper-level groups, and the ability to exploit partial information in prediction when the presence or resolution of observed taxa differs between two datasets. While the tree-based fused lasso provides more flexibility with regard to level of association, by aggregating along branches of a tree, differing associations within a group cannot be combined with the inclusion of upper-level features in the model [130]. In addition, for prediction on new datasets, all existing methods assume the set of observed taxa is the same between the model-building dataset and the prediction set. In contrast, MCL is able to include group-level information contributed by a taxon even if there are mismatches in the lower-level (e.g., genus-level) feature set. Despite this added flexibility, the usual restrictions in prediction performance also apply when using MCL. Demographic, behavioral, and clinical differences between the model-building data and the prediction data will influence the accuracy of predictions.

Simulations show that, when the features in the prediction set match those in the model-building dataset, MCL has prediction error similar to or lower than CL when the data are generated from the multilevel model, or when data are generated from the linear log-contrast model but nonzero coefficients at least partially follow the presumed grouping structure. When some taxa are identified only to the group level in the prediction set, prediction error

is lower with MCL than CL in all settings.

In the American Gut Project data, we find both class-level and genus-level associations between the gut microbiome and BMI. Our results suggest that Actinobacteria are positively associated with BMI at the class level, whereas associations with genera belonging to the Firmicutes classes, Bacilli and Clostridia, tend to vary. Prediction error in new data, both from the American Gut Project and the Human Microbiome Project, is similar for MCL as for CL applied at class level. This is consistent with the small proportion of non-zero within-group features; MCL is expected to outperform CL at the class level when nonzero features are more evenly balanced between group-level and within-group features. In both new datasets, prediction error was substantially higher for CL-genus level.

MCL is conceptually similar to the Mixed effects Score Test (MiST) for rare genetic variants [118]. In that setting, because the variants are rare, power for testing individual variants is limited. MiST uses a hierarchical linear model to encode which uses a hierarchical model to specify fixed effects corresponding to variant characteristics (e.g., nonsense or missense) and variant-specific random effects. The analogous method for the microbiome would borrow information across taxa within a group using taxon-specific random effects instead of MCL's within-group fixed effects.

The choice of which two phylogenetic levels to include in an MCL model is not always clear, and extension to more than two levels could further improve flexibility in new data and prediction accuracy. For non-compositional data with rare features, [138] proposes aggregating features along a tree and assigning a coefficient to each branch of the tree. A fusion penalty within a tree encourages nodes attached to the same branch to have the same coefficient, which amounts to including only the parent node (similar to our group-level effect). However, this approach does not account for the compositionality or subcompositionality of microbiome data. A similar procedure may be productive for extending MCL to more than two levels.

Overall, the multilevel linear log-contrast model provides flexibility in the face of unknown level of association, potentially differing associations within groups of taxa, and the ability

to exploit partial information in prediction when the presence or resolution of observed taxa differs between two datasets. Inducing sparsity at the group level and within groups further improves prediction error and interpretability; in addition to identifying associated taxa, MCL results give insight into taxonomic levels of association. As microbiome profiles are increasingly used for diagnosis and prognosis, MCL can provide a framework for phenotype prediction using compositional data with potentially differing feature sets.

Table 5.4: Nonzero genus-level coefficients for MCL and CL-genus. n/a indicates a singleton group, for which no within-group effect is estimated in the MCL model.

Phylum	Class	Genus	MCL $\hat{\beta}$	MCL $\hat{\gamma}$	CL $\hat{\beta}$
Actinobacteria	Actinobacteria	Brevibacterium	0.251	·	-0.152
Actinobacteria	Actinobacteria	Corynebacterium	0.251	-0.039	·
Actinobacteria	Actinobacteria	Rhodococcus	0.251	·	1.125
Actinobacteria	Actinobacteria	Bifidobacterium	0.251	0.046	0.081
Actinobacteria	Actinobacteria	Unclassified Bifidobacteriaceae	0.251	·	0.339
Bacteroidetes	Bacteroidia	Alistipes	·	·	0.160
Bacteroidetes	Flavobacteriia	Unclassified Weeksellaceae	·	·	-0.157
Cyanobacteria	4C0d-2	Unclassified YS2	-0.181	n/a	-0.201
Cyanobacteria	Chloroplast	Unclassified Streptophyta	-0.214	n/a	-0.254
Firmicutes	Bacilli	Unclassified Bacillaceae	·	0.039	·
Firmicutes	Bacilli	Enterococcus	·	-0.104	-0.109
Firmicutes	Bacilli	Streptococcus	·	0.085	0.040
Firmicutes	Clostridia	Unclassified Mogibacteriaceae	·	-0.182	-0.095
Firmicutes	Clostridia	Peptoniphilus	·	·	0.028
Firmicutes	Clostridia	SMB53	·	-0.011	·
Firmicutes	Clostridia	Coprococcus	0	-0.052	·
Firmicutes	Clostridia	Dorea	·	0.046	·
Firmicutes	Clostridia	Unclassified Lachnospiraceae	·	0.058	·
Firmicutes	Clostridia	Unclassified Clostridiales	·	-0.116	·
Firmicutes	Clostridia	rc4-4 (Family Peptococcaceae)	·	0.296	0.365
Firmicutes	Clostridia	Faecalibacterium	·	-0.025	·
Firmicutes	Clostridia	Oscillospira	·	-0.024	-0.056
Firmicutes	Erysipelotrichi	[Clostridium]	·	·	-0.183
Lentisphaerae	[Lentisphaeria]	Unclassified Victivallaceae	-0.153	·	-0.050
Lentisphaerae	[Lentisphaeria]	Victivallis	-0.153	·	-0.183
Proteobacteria	Alphaproteobacteria	Unclassified Rhizobiales	·	·	-0.884
Proteobacteria	Alphaproteobacteria	Sphingopyxis	·	·	-0.010
Proteobacteria	Betaproteobacteria	Oxalobacter	-0.122	·	-0.125
Proteobacteria	Deltaproteobacteria	Bilophila	0.346	·	0.126
Proteobacteria	Gammaproteobacteria	Citrobacter	·	·	0.081
Proteobacteria	Gammaproteobacteria	Aggregatibacter	·	·	-0.046
Proteobacteria	Gammaproteobacteria	Haemophilus	·	·	-0.083
Proteobacteria	Gammaproteobacteria	Unclassifeid Moraxellaceae	·	·	0.339
Synergistetes	Synergistia	Unclassified Synergistaceae	·	·	-0.058
Tenericutes	Mollicutes	Unclassified Anaeroplasmataceae	-0.015	·	-0.003
Tenericutes	RF3	Unclassified	-0.019	n/a	-0.036

Chapter 6

CONCLUSIONS AND FUTURE WORK

Although investigation of the microbiome has become a valuable part of scientific and clinical investigation into health, disease, and treatment strategies, new methods for microbiome analysis are needed to accommodate modern study designs and to incorporate structural information into identification of associated taxa and predictions based on the microbiome. The methods proposed in this dissertation add valuable tools to a microbiologist's or bioinformatician's toolbox in both of these areas. They also point towards new avenues for further statistical methods development.

One potential direction for future work is performing feature selection within distance-based analyses. This dual approach would retain the power advantages of distance-based analysis, while leveraging the interpretability of feature selection models. Most distance metrics, including the UniFrac family of distances and dissimilarities, are essentially weighted linear combinations of taxa. Therefore, the taxon-specific terms could be multiplied by an additional, penalized weight vector in the kernel definition. Taxa with non-zero weights are those whose contribution to the dissimilarity between communities is most important.

Another important area of further development is to incorporate more than two levels of phylogenetic grouping in the penalized regression models. This could be done, for example, by adapting the approach of [138] to the compositional data setting, assigning coefficients to branches of a phylogenetic tree rather than nodes. It would also be valuable to allow outcome types beyond simple quantitative phenotypes, since classification tasks and time-to-event outcomes are of particular interest as microbiome profiling is incorporated into clinical studies. Extensions of penalized linear regression to other outcome types are well developed for non-compositional data, and similar changes to the loss function will extend

CSGL and MCL to studies with binary or censored time-to-event outcomes.

There are additional challenges associated with microbiome data that we do not consider. In particular, while zero taxon counts are often cited as a problem, it is unclear how much influence the choice of zero-adjustment method has on the results of an analysis, and few statistically informed zero-adjustment approaches exist. A robust comparison of the impact of existing methods to handle excess zeros would be an extremely useful resource for establishing best practices in real microbiome data analysis. Also, taxon assignment is often based on short sequencing reads that are compared to standard databases. The quality of these databases underlies the accuracy of phylogenetic relationships among taxa, and uncertainties in phylogenetic assignments are rarely accounted for. While these analysis steps are external to the methods presented herein, they influence the accuracy and scientific relevance of analysis results and are broadly useful questions that could direct future work in statistical microbiome analysis.

BIBLIOGRAPHY

[1] J Aitchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[2] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[3] John Aitchison and John Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984.

[4] Michael Altenbuchinger, Thorsten Rehberg, HU Zacharias, Frank Stämmler, Katja Dettmer, Daniela Weber, Andreas Hiergeist, Andre Gessner, Ernst Holler, Peter J Oefner, and Rainer Spang. Reference point insensitive molecular data analysis. *Bioinformatics*, 33(2):219–226, 2016.

[5] Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Justine Debelius, James T Morton, Embriette Hyde, Adam Robbins-Pianka, and Rob Knight. Correcting for microbial blooms in fecal samples during room-temperature shipping. *MSystems*, 2(2):e00199–16, 2017.

[6] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001.

[7] Fabrice Armougom, Mireille Henry, Bernard Vialettes, Denis Raccah, and Didier Raoult. Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *PloS ONE*, 4(9):e7125, 2009.

[8] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[9] Francesco Asnicar, George Weingart, Timothy L Tickle, Curtis Huttenhower, and Nicola Segata. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 3:e1029, 2015.

[10] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(Jun):1179–1225, 2008.

[11] Hans Bisgaard, Nan Li, Klaus Bonnelykke, Bo Lund Krogsgaard Chawes, Thomas Skov, Georg Paludan-Müller, Jakob Stokholm, Birgitte Smith, and Karen Angeliki Krogfelt. Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *Journal of Allergy and Clinical Immunology*, 128(3):646–652, 2011.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[13] J Roger Bray and John T Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[14] Hilary P Browne, Samuel C Forster, Blessing O Anonye, Nitin Kumar, B Anne Neville, Mark D Stares, David Goulding, and Trevor D Lawley. Culturing of unculturablehuman microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543, 2016.

[15] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science and Business Media, 2011.

[16] Tianxi Cai, Giulia Tonini, and Xihong Lin. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–986, 2011.

[17] Soumitesh Chakravorty, Danica Helb, Michele Burday, Nancy Connell, and David Alland. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, 69(2):330–339, 2007.

[18] Qin Chang, Yihui Luan, and Fengzhu Sun. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, 12(1):1, 2011.

[19] Emily S Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D Bushman, and Ronald G Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*, 5(12):e15216, 2010.

[20] Han Chen, Thomas Lumley, Jennifer Brody, Nancy L Heard-Costa, Caroline S Fox, L Adrienne Cupples, and Josée Dupuis. Sequence kernel association test for survival traits. *Genetic Epidemiology*, 38(3):191–197, 2014.

[21] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[22] Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2012.

[23] Jun Chen, Wenan Chen, Ni Zhao, Michael C Wu, and Daniel J Schaid. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic Epidemiology*, 40(1):5–19, 2016.

[24] Jun Chen and Hongzhe Li. Kernel methods for regression analysis of microbiome compositional data. In *Topics in Applied Statistics*, pages 191–201. Springer, 2013.

[25] Bruno P Chumpitazi, Julia L Cope, Emily B Hollister, Cynthia M Tsai, Ann R McMeans, Ruth A Luna, James Versalovic, and Robert J Shulman. Randomised clinical trial: gut microbiome biomarkers are associated with clinical response to a low FODMAP diet in children with the irritable bowel syndrome. *Alimentary Pharmacology & Therapeutics*, 42(4):418–427, 2015.

[26] Siobhan F Clarke, Eileen F Murphy, Kanishka Nilaweera, Paul R Ross, Fergus Shanahan, Paul W OToole, and Paul D Cotter. The gut microbiota and its relationship to diet and obesity: new insights. *Gut Microbes*, 3(3):186–202, 2012.

[27] Thomas Clavel, Charles Desmarchelier, Dirk Haller, Philippe Gérard, Sascha Rohn, Patricia Lepage, and Hannelore Daniel. Intestinal microbiota in metabolic diseases: from bacterial community structure and functions to species of pathophysiological relevance. *Gut Microbes*, 5(4):544–551, 2014.

[28] Jose C Clemente, Luke K Ursell, Laura Wegener Parfrey, and Rob Knight. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270, 2012.

[29] James R Cole, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1):D633–D642, 2013.

[30] David R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[31] R. B. Davies. The distribution of a linear combination of chi-2 random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 29(3):323–333, 1980.

[32] Tristano Bacchetti De Gregoris, Nick Aldred, Anthony S Clare, and J Grant Burgess. Improvement of phylum-and class-specific primers for real-time pcr quantification of bacterial taxa. *Journal of Microbiological Methods*, 86(3):351–356, 2011.

[33] Tim GJ de Meij, Evelien FJ de Groot, Anat Eck, Andries E Budding, CM Frank Kneepkens, Marc A Benninga, Adriaan A van Bodegraven, and Paul HM Savelkoul. Characterization of microbiota in children with chronic functional constipation. *PloS ONE*, 11(10):e0164731, 2016.

[34] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006.

[35] GG Donders, Kristel Van Calsteren, G Bellen, R Reybrouck, T Van den Bosch, I Riphagen, and S Van Lierde. Predictive value for preterm birth of abnormal vaginal flora, bacterial vaginosis and aerobic vaginitis during the first trimester of pregnancy. *BJOG: An International Journal of Obstetrics & Gynaecology*, 116(10):1315–1324, 2009.

[36] Sylvia H Duncan, GE Lobley, Grietje Holtrop, J Ince, AM Johnstone, Petra Louis, and Harry James Flint. Human colonic microbiota associated with diet, obesity and weight loss. *International Journal of Obesity*, 32(11):1720, 2008.

[37] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

[38] John R Erb-Downward, Deborah L Thompson, Meilan K Han, Christine M Freeman, Lisa McCloskey, Lindsay A Schmidt, Vincent B Young, Galen B Toews, Jeffrey L Curtis, Baskaran Sundaram, et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS ONE*, 6(2):e16384, 2011.

[39] Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

[40] Michael A Fischbach and Julia A Segre. Signaling in host-associated microbial communities. *Cell*, 164(6):1288–1300, 2016.

[41] Jingyuan Fu, Marc Jan Bonder, María Carmen Cenit, Ettje F Tigchelaar, Astrid Maatman, Jackie AM Dekens, Eelke Brandsma, Joanna Marczynska, Floris Imhann, Rinse K Weersma, et al. The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circulation Research*, 117(9):817–824, 2015.

[42] Brian R Gaines and Hua Zhou. Algorithms for fitting the constrained lasso. *arXiv preprint arXiv:1611.01511*, 2016.

[43] Philippe Gérard. Gut microbiota and obesity. *Cellular and Molecular Life Sciences*, 73(1):147–162, 2016.

[44] Jonathan L Golob, Steven A Pergam, Sujatha Srinivasan, Tina L Fiedler, Congzhou Liu, Kristina Garcia, Marco Mielcarek, Daisy Ko, Sarah Aker, Sara Marquis, Tillie Loeffelholz, Anna Plantinga, Michael C Wu, Kevin Celustka, Alex Morrison, Maresa Woodfield, and David N Fredricks. Stool microbiota at neutrophil recovery is predictive for severe acute graft vs host disease after hematopoietic cell transplantation. *Clinical Infectious Diseases*, 65(12):1984–1991, 2017.

[45] Julia K Goodrich, Jillian L Waters, Angela C Poole, Jessica L Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T Bell, et al. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, 2014.

[46] J Graessler, Y Qin, H Zhong, J Zhang, Julio Licinio, Ma-Li Wong, A Xu, T Chavakis, AB Bornstein, Monika Ehrhart-Bornstein, et al. Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. *The Pharmacogenomics Journal*, 13(6):514, 2013.

[47] MeiLan K Han, Yueren Zhou, Susan Murray, Nabihah Tayob, Imre Noth, Vibha N Lama, Bethany B Moore, Eric S White, Kevin R Flaherty, Gary B Huffnagle, et al. Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *The Lancet Respiratory Medicine*, 2(7):548–556, 2014.

[48] Irva Hertz-Picciotto and Beverly Rockhill. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, pages 1151–1156, 1997.

[49] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[50] Shernan G Holtan, Todd E DeFor, Aleksandr Lazaryan, Nelli Bejanyan, Mukta Arora, Claudio G Brunstein, Bruce R Blazar, Margaret L MacMillan, and Daniel J Weisdorf. Composite end point of graft-versus-host disease-free, relapse-free survival after allogeneic hematopoietic cell transplantation. *Blood*, 125(8):1333–1338, 2015.

[51] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[52] Zehra Esra Ilhan, John K DiBaise, Nancy G Isern, David W Hoyt, Andrew K Marcus, Dae-Wook Kang, Michael D Crowell, Bruce E Rittmann, and Rosa Krajmalnik-Brown. Distinctive microbiomes and metabolites linked with weight loss after gastric bypass, but not gastric banding. *The ISME Journal*, 11(9):2047, 2017.

[53] Gareth M James, Courtney Paulson, and Paat Rusmevichientong. The constrained lasso. In *Refereed Conference Proceedings*, volume 31, pages 4945–4950, 2012.

[54] Ian B Jeffery, Marcus J Claesson, Paul W O'toole, and Fergus Shanahan. Categorization of the gut microbiota: enterotypes or gradients? *Nature Reviews Microbiology*, 10(9):591, 2012.

[55] Robert R Jenq, Ying Taur, Sean M Devlin, Doris M Ponce, Jenna D Goldberg, Katya F Ahr, Eric R Littmann, Lilan Ling, Asia C Gobourne, Liza C Miller, et al. Intestinal blautia is associated with reduced death from graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 21(8):1373–1383, 2015.

[56] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, and Gane K-S Wong. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7:459, 2016.

[57] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

[58] George S Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[59] Dan Knights, Laura Wegener Parfrey, Jesse Zaneveld, Catherine Lozupone, and Rob Knight. Human-associated microbial signatures: examining their predictive value. *Cell Host & Microbe*, 10(4):292–296, 2011.

[60] Dan Knights, Tonya L Ward, Christopher E McKinlay, Hannah Miller, Antonio Gonzalez, Daniel McDonald, and Rob Knight. Rethinking "enterotypes". *Cell Host & Microbe*, 16(4):433–437, 2014.

[61] Hyunwook Koh, Martin J Blaser, and Huilin Li. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*, 5(1):45, 2017.

[62] Hyunwook Koh, Alexandra E Livanos, Martin J Blaser, and Huilin Li. A highly adaptive microbiome-based association test for survival traits. *BMC Genomics*, 19(1):210, 2018.

[63] Omry Koren, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E Ley. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, 9(1):e1002863, 2013.

[64] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, José C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47, 2012.

[65] Purnima S Kumar, Michael R Brooker, Scot E Dowd, and Terry Camerlengo. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PloS ONE*, 6(6):e20956, 2011.

[66] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5):e1004226, 2015.

[67] Jean-Christophe Lagier, Saber Khelaifia, Maryam Tidjani Alou, Sokhna Ndongo, Niokhor Dione, Perrine Hugon, Aurelia Caputo, Frederic Cadoret, Sory Ibrahima Traore, Gregory Dubourg, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology*, 1(12):16203, 2016.

[68] Sophie Lambert-Lacroix, Frédérique Letué, et al. Partial least squares and Cox model with application to gene expression. *Technical report*, 2011.

[69] Stacey M Lambeth, Trechelle Carson, Janae Lowe, Thiruvarangan Ramaraj, Jonathan W Leff, Li Luo, Callum J Bell, and Vallabh O Shah. Composition, diversity and abundance of gut microbiome in prediabetes and type 2 diabetes. *Journal of Diabetes and Obesity*, 2(3):1, 2015.

[70] Jennifer T Lau, Fiona J Whelan, Isiri Herath, Christine H Lee, Stephen M Collins, Premysl Bercik, and Michael G Surette. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome medicine*, 8(1):72, 2016.

[71] Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–11075, 2005.

[72] Ruth E Ley, Daniel A Peterson, and Jeffrey I Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–848, 2006.

[73] Hong Lin, Yanpeng An, Fuhua Hao, Yulan Wang, and Huiru Tang. Correlations of fecal metabonomic and microbiomic changes induced by high-fat diet in the pre-obesity state. *Scientific Reports*, 6:21618, 2016.

[74] Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, page asu031, 2014.

[75] Xinyi Lin, Tianxi Cai, Michael C Wu, Qian Zhou, Geoffrey Liu, David C Christiani, and Xihong Lin. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, 35(7):620–631, 2011.

[76] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

[77] Zongzhi Liu, Todd Z DeSantis, Gary L Andersen, and Rob Knight. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, 36(18):e120–e120, 2008.

[78] Rohit Loomba, Victor Seguritan, Weizhong Li, Tao Long, Niels Klitgord, Archana Bhatt, Parambir Singh Dulai, Cyrielle Caussy, Richele Bettencourt, Sarah K Highlander, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metabolism*, 25(5):1054–1062, 2017.

[79] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[80] Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

[81] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26, 2015.

[82] Chaysavanh Manichanh, Lionel Rigottier-Gois, Elian Bonnaud, Karine Gloux, Eric Pelletier, Lionel Frangeul, Renaud Nalin, Cyrille Jarrin, Patrick Chardon, Phillipe Marteau, et al. Reduced diversity of faecal microbiota in crohns disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, 2006.

[83] Josep-Antoni Martín-Fernández, Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

[84] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

[85] Conor J Meehan and Robert G Beiko. A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biology and Evolution*, 6(3):703–713, 2014.

[86] M Million, J-C Lagier, D Yahav, and M Paul. Gut bacterial microbiota and obesity. *Clinical Microbiology and Infection*, 19(4):305–313, 2013.

[87] Caroline M Mitchell, Sujatha Srinivasan, Anna Plantinga, Michael C Wu, Susan D Reed, Katherine A Guthrie, Andrea Z LaCroix, Tina Fiedler, Matthew Munch, Congzhou Liu, Noah G Hoffman, Ian A Blair, Katherine Newton, Ellen W Freeman, Hadine W Joffe, Lee Cohen, and David N Fredricks. Associations between improvement in genitourinary symptoms of menopause and changes in the vaginal ecosystem. *Menopause*, 25(5):500–507, 2018.

[88] Shirajum Monira, Shota Nakamura, Kazuyoshi Gotoh, Kaori Izutsu, Haruo Watanabe, Nur Haque Alam, Hubert Ph Endtz, Alejandro Cravioto, Sk Ali, Takaaki Nakaya, et al. Gut microbiota of healthy and malnourished children in Bangladesh. *Frontiers in Microbiology*, 2:228, 2011.

[89] Ember M Morrissey, Rebecca L Mau, Egbert Schwartz, J Gregory Caporaso, Paul Dijkstra, Natasja van Gestel, Benjamin J Koch, Cindy M Liu, Michaela Hayer, Theresa A McHugh, Jane C Marks, Lance B Price, and Bruce A Hungate. Phylogenetic organization of bacterial activity. *The ISME Journal*, 10(9):2336, 2016.

[90] Brian D Muegge, Justin Kuczynski, Dan Knights, Jose C Clemente, Antonio González, Luigi Fontana, Bernard Henrissat, Rob Knight, and Jeffrey I Gordon. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, 332(6032):970–974, 2011.

[91] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.

[92] I Nadal, A Santacruz, A Marcos, J Warnberg, M Garagorri, LA Moreno, M Martin-Matillas, C Campoy, A Martí, A Moleres, et al. Shifts in clostridia, bacteroides and immunoglobulin-coating fecal bacteria associated with weight loss in obese adolescents. *International Journal of Obesity*, 33(7):758, 2009.

[93] Ståle Nygård, Ørnulf Borgan, Ole Christian Lingjærde, and Hege Leite Størvold. Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, 14(2):179–195, 2008.

[94] Andrew B Onderdonk, Mary L Delaney, and Raina N Fichorova. The human microbiome during bacterial vaginosis. *Clinical Microbiology Reviews*, 29(2):223–238, 2016.

[95] Peter J Park, Lu Tian, and Isaac S Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18(suppl 1):S120–S127, 2002.

[96] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200, 2013.

[97] John Penders, Kerstin Gerhold, Ellen E Stobberingh, Carel Thijs, Kurt Zimmermann, Susanne Lau, and Eckard Hamelmann. Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood. *Journal of Allergy and Clinical Immunology*, 132(3):601–607, 2013.

[98] Laurent Philippot, Siv GE Andersson, Tom J Battin, James I Prosser, Joshua P Schimel, William B Whitman, and Sara Hallin. The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology*, 8(7):523, 2010.

[99] Anna Plantinga, Xiang Zhan, Ni Zhao, Jun Chen, Robert R Jenq, and Michael C Wu. MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome*, 5(1):17, 2017.

[100] Geoffrey A Preidis and James Versalovic. Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era. *Gastroenterology*, 136(6):2015–2031, 2009.

[101] Ruben Props, Frederiek-Maarten Kerckhof, Peter Rubbens, Jo De Vrieze, Emma Hernandez Sanabria, Willem Waegeman, Pieter Monsieurs, Frederik Hammes, and Nico Boon. Absolute quantification of microbial taxon abundances. *The ISME Journal*, 11(2):584, 2017.

[102] Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196, 2007.

[103] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59, 2010.

[104] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[105] Timothy W Randolph, Sen Zhao, Wade Copeland, Meredith Hullar, Ali Shojaie, et al. Kernel-penalized regression for analysis of microbiome data. *The Annals of Applied Statistics*, 12(1):540–566, 2018.

[106] Alessandra Riva, Francesca Borgo, Carlotta Lassandro, Elvira Verduci, Giulia Morace, Elisa Borghi, and David Berry. Pediatric obesity is associated with an altered gut microbiota and discordant shifts in Firmicutes populations. *Environmental Microbiology*, 19(1):95–105, 2017.

[107] Bertrand Routy, Emmanuelle Le Chatelier, Lisa Derosa, Connie PM Duong, Maryam Tidjani Alou, Romain Daillère, Aurélie Fluckiger, Meriem Messaoudene, Conrad Rauber, Maria P Roberti, et al. Gut microbiome influences efficacy of pd-1–based immunotherapy against epithelial tumors. *Science*, 359(6371):91–97, 2018.

[108] Patrick D Schloss, Dirk Gevers, and Sarah L Westcott. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, 6(12):e27310, 2011.

[109] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[110] Andreas Schwiertz, David Taras, Klaus Schäfer, Silvia Beijer, Nicolaas A Bos, Christiane Donus, and Philip D Hardt. Microbiota and SCFA in lean and overweight healthy subjects. *Obesity*, 18(1):190–195, 2010.

[111] Tanusree Sen, Carolina R Cawthon, Benjamin Thomas Ihde, Andras Hajnal, Patricia M DiLorenzo, B Claire, and Krzysztof Czaja. Diet-driven microbiota dysbiosis is associated with vagal remodeling and obesity. *Physiology & Behavior*, 173:305–317, 2017.

[112] Ron Sender, Shai Fuchs, and Ron Milo. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–340, 2016.

[113] Pixu Shi, Anru Zhang, Hongzhe Li, et al. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016.

[114] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[115] Sujatha Srinivasan, Noah G Hoffman, Martin T Morgan, Frederick A Matsen, Tina L Fiedler, Robert W Hall, Frederick J Ross, Connor O McCoy, Roger Bumgarner, Jeanne M Marrazzo, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PloS ONE*, 7(6):e37818, 2012.

[116] E Stackebrandt and BM Goebel. Taxonomic note: a place for dna-dna reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, 44(4):846–849, 1994.

[117] George PH Styan. Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*, 57(3):567–572, 1970.

[118] Jianping Sun, Yingye Zheng, and Li Hsu. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, 37(4):334–344, 2013.

[119] Olga Tanaseichuk, James Borneman, and Tao Jiang. Phylogeny-based classification of microbial communities. *Bioinformatics*, 30(4):449–456, 2013.

[120] Zheng-Zheng Tang, Guanhua Chen, and Alexander V Alekseyenko. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, 32(17):2618–2625, 2016.

[121] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, pages 267–288, 1996.

[122] Ryan Joseph Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.

[123] Herbert Tilg and Arthur Kaser. Gut microbiome, obesity, and metabolic dysfunction. *The Journal of Clinical Investigation*, 121(6):2126–2132, 2011.

[124] Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5):330–335, 2016.

[125] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[126] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027, 2006.

[127] Peter J Turnbaugh, Vanessa K Ridaura, Jeremiah J Faith, Federico E Rey, Rob Knight, and Jeffrey I Gordon. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*, 1(6):6ra14–6ra14, 2009.

[128] Alan W Walker, Jennifer C Martin, Paul Scott, Julian Parkhill, Harry J Flint, and Karen P Scott. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*, 3(1):26, 2015.

[129] Lin Wang, Peicheng Li, Zhaosheng Tang, Xinfeng Yan, and Bo Feng. Structural modulation of the gut microbiota and the relationship with body weight: compared evaluation of liraglutide and saxagliptin treatment. *Scientific Reports*, 6:33251, 2016.

[130] Tao Wang and Hongyu Zhao. Constructing predictive microbial signatures at multiple taxonomic levels. *Journal of the American Statistical Association*, 112(519):1022–1031, 2017.

[131] Tao Wang, Hongyu Zhao, et al. Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11(2):771–791, 2017.

[132] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[133] Daniela Weber, Peter J Oefner, Andreas Hiergeist, Josef Koestler, André Gessner, Markus Weber, Joachim Hahn, Daniel Wolff, Frank Stämmler, Rainer Spang, Wolfgang Herr, Katja Dettmer, and Ernst Holler. Low urinary indoxyl sulfate levels early after transplantation reflect a disrupted microbiome and are associated with poor outcome. *Blood*, 126(14):1723–1728, 2015.

[134] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 2017.

[135] James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Computational Biology*, 5(4):e1000352, 2009.

[136] Chong Wu, Jun Chen, Junghi Kim, and Wei Pan. An adaptive association test for microbiome data. *Genome Medicine*, 8(1):56, 2016.

[137] Yujia Wu, Xiaopei Chi, Qian Zhang, Feng Chen, and Xuliang Deng. Characterization of the salivary microbiome in people with obesity. *PeerJ*, 6:e4458, 2018.

[138] Xiaohan Yan and Jacob Bien. Rare feature selection in high dimensions. *arXiv preprint arXiv:1803.06675*, 2018.

[139] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222, 2012.

[140] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67, 2006.

[141] Georg Zeller, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766, 2014.

[142] Jing Zhai, Juhyun Kim, Kenneth S Knox, Homer L Twigg III, Hua Zhou, and Jin J Zhou. Variance component selection with applications to microbiome taxonomic data. *Frontiers in Microbiology*, 9:509, 2018.

[143] Jing Zhai, Kenneth S Knox, Homer L Twigg, Hua Zhou, and Jin J Zhou. Exact tests of zero variance component in presence of multiple variance components with application to longitudinal microbiome study. *bioRxiv*, page 281246, 2018.

[144] Xiang Zhan, Xingwei Tong, Ni Zhao, Arnab Maity, Michael C Wu, and Jun Chen. A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 41(3):210–220, 2017.

[145] Husen Zhang, John K DiBaise, Andrea Zuccolo, Dave Kudrna, Michele Braidotti, Yeisoo Yu, Prathap Parameswaran, Michael D Crowell, Rod Wing, Bruce E Rittmann, et al. Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, 106(7):2365–2370, 2009.

[146] Xiuying Zhang, Dongqian Shen, Zhiwei Fang, Zhuye Jie, Xinmin Qiu, Chunfang Zhang, Yingli Chen, and Linong Ji. Human gut microbiota changes reveal the progression of glucose intolerance. *PloS ONE*, 8(8):e71108, 2013.

[147] Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807, 2015.

[148] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Iteratively Reweighted Least Squares Algorithm for Cox Model

An iteratively reweighted least squares (IRLS) algorithm can be used to fit the linear model at convergence that is equivalent to the Cox PH model of interest. At the kth step of the IRLS algorithm, we solve

$$\tilde{y}^k = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2(\tilde{W}^k)^{-1})$$

with weight matrix

$$\tilde{W}^k = \text{diag}\left(\int_0^\infty I(T_i \geq t)e^{X_i'\tilde{\beta}^k} d\hat{\Lambda}_0(t) - \int_0^\infty I(T_i \geq t)w_i(\tilde{\beta}^k, t)e^{X_i'\tilde{\beta}^k} d\hat{\Lambda}_0(t)\right)$$

where $w_i(t) = \frac{e^{X_i^\top \beta}}{\sum_{l=1}^n Y_l(t)e^{X_l'\beta}}$, and working response

$$\tilde{y} = X\tilde{\beta}^{k-1} + (\tilde{W}^{k-1})^{-1}\hat{\mathbf{M}}^{k-1}.$$

The corresponding quantities without the superscript k refer to the model at convergence. Then the modified score statistic is equivalent to

$$Q^* = \frac{(\tilde{y} - X\tilde{\beta})'\tilde{W}\mathbf{K}\tilde{W}(\tilde{y} - X\tilde{\beta})}{\hat{\sigma}^2}$$

which is analogous to the linear and logistic cases considered in [23]. Multiplying both sides of the equation by $\tilde{W}^{1/2}$ and defining $\tilde{y}^* = W^{1/2}\tilde{y}$, $X^* = W^{1/2}X$, and $\epsilon^* = W^{1/2}\epsilon$, the model can be expressed as

$$\tilde{y}^* = X^*\beta + \epsilon^*, \ \epsilon^* \sim \mathcal{N}(0, \sigma^2 I)$$

with projection matrix $P_0^* = I - X^*(X^{*'}X^*)^{-1}X^{*'}$.

A.2 Relationship Between Estimated and True Residuals

To derive the relationship between $\hat{M}$ and M, recall that

$$\begin{aligned} \hat{z} &= X\hat{\beta} + W^{-1}\hat{M} \\ z &= X\beta + W^{-1}M \\ \hat{\beta} &= (X'WX)^{-1}X'Wz \end{aligned}$$

Then, solving the first equation for $\hat{M}$ gives

$$\begin{aligned} \hat{M} &= W(z - X\hat{\beta}) = W\left[X\beta + W^{-1}M - X\hat{\beta}\right] \\ &= W\left[I - X(X'WX)^{-1}X'W\right]z \\ &= W\left[I - X(X'WX)^{-1}X'W\right](X\beta + W^{-1}M) \\ &= \left[WW^{-1}M - WX(X'WX)^{-1}X'WW^{-1}M\right] \\ &= \left[I - WX(X'WX)^{-1}X'\right]M \\ &= \left[I - X^*(X^{*'}X^*)^{-1}X^*\right]M = P_0^*M \end{aligned}$$

so that $\hat{M} = P_0^*M$, as claimed.

A.3 Additional Simulation Results

Table A.1: Empirical Type I errors when no small sample correction is used, with approximately 25% censoring. Results are based on 5,000 simulated datasets. K_u, $K_{0.5}$, K_w, and K_{BC} represent results for the unweighted UniFrac kernel, generalized UniFrac kernel with $\alpha = 0.5$, weighted UniFrac kernel, and Bray-Curtis kernel, respectively.

n	K_u	$K_{0.5}$	K_w	K_{BC}
100	0.0102	0.0084	0.0286	0.0144
200	0.0162	0.0130	0.0308	0.0188
500	0.0218	0.0248	0.0364	0.0254

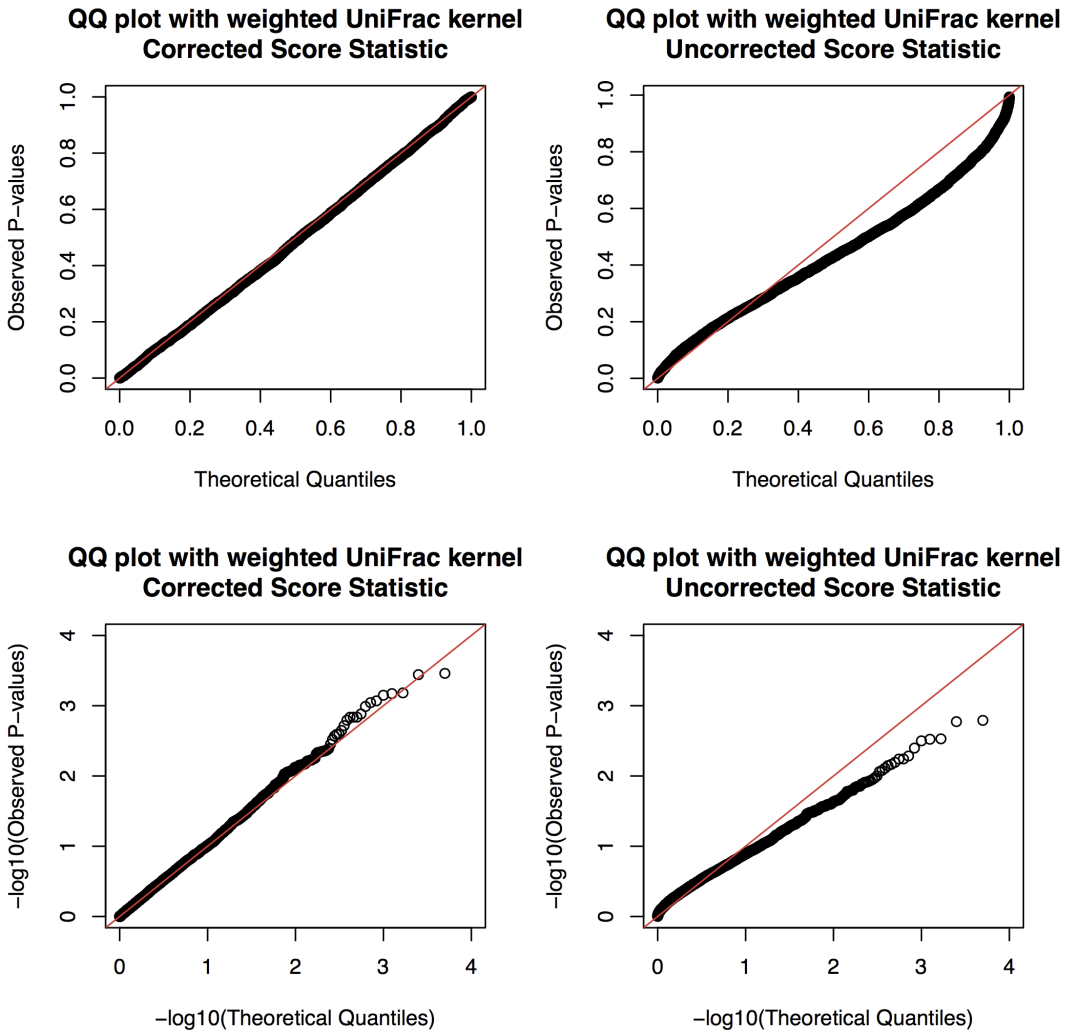


Figure A.1: QQ plots demonstrating the performance of the uncorrected and corrected score statistics. The lower two panels display the -$\log_{10}$(P-values) for both statistics plotted against the theoretical distribution, showing that MiRKAT-S with the correction behaves as expected, whereas without the correction, the p-values deviate significantly from the theoretical distribution. The upper two panels provide scientific insight regarding p-values generated using the uncorrected statistic. In particular, true p-values lower than 0.2 will tend to be too large when using the uncorrected statistic, whereas true p-values higher than 0.2 will tend to be too small when using the uncorrected statistic.

Appendix B

APPENDIX TO CHAPTER 3

B.1 Unweighted LUniFrac is a Distance

To prove that the unweighted LUniFrac measure is a true distance metric, we need to show that:

1. $D_{XY} \geq 0$ (nonnegativity)

2. $D_{XY} = 0$ iff $X = Y$ (identity of indiscernables)

3. $D_{XY} = D_{YX}$ (symmetry)

4. $D_{XZ} \leq D_{XY} + D_{YZ}$ (triangle inequality)

Nonnegativity: By definition, branch lengths satisfy $b_i > 0$. The remaining term in the metric is an absolute value term ($|d_i^X - d_i^Y| \geq 0$). Therefore, $D_{XY} \geq 0$.
Identity: In order to demonstrate this item, we must first define what we mean by "indiscernables."

Let X and Y be two microbial communities measured at two time points each, so we have $X(t_1)$ and $X(t_2)$, and similarly, $Y(t_1)$ and $Y(t_2)$. The mapping from $X_i(t_1)$ and $X_i(t_2)$ to d_i^X is not 1-1. However, the quantity of interest for this metric is not the microbial community itself, it is changes in microbial communities. Therefore, in this case the relevant "indiscernables" are d^X and d^Y rather than X and Y. That is, we must demonstrate that $D_{XY} = 0$ iff $d^X = d^Y$, regardless of the original communities $X(t_1)$, $X(t_2)$, $Y(t_1)$, and $Y(t_2)$,

Suppose $d_i^X = d_i^Y$ for all taxa i. Then $d_i^X - d_i^Y = 0 \, \forall i$, and so $\sum_{i=1}^{p} b_i |d_i^X - d_i^Y|/2 = 0$. Therefore $X = Y \implies D_{XY} = 0$.

Now suppose $D_{XY} = 0$. Because each term is nonnegative and branch lengths b_i are strictly positive ($b_i > 0\,\forall i$), this implies that $|d_i^X - d_i^Y| = 0\,\forall i$. This only holds if $d_i^X = d_i^Y\ \forall i$. Therefore $D_{XY} = 0 \implies d^X = d^Y$.

Symmetry: Here, we note that

$$\begin{aligned} D_{XY} &= \frac{\sum_{i=1}^{p} b_i |d_i^X - d_i^Y|/2}{\sum_{i=1}^{p} b_i} = \frac{\sum_{i=1}^{p} b_i |(-1) \times (d_i^X - d_i^Y)|/2}{\sum_{i=1}^{p} b_i} \\ &= \frac{\sum_{i=1}^{p} b_i |d_i^Y - d_i^X|/2}{\sum_{i=1}^{p} b_i} = D_{YX} \end{aligned}$$

Triangle Inequality: Using the triangle inequality for the absolute value ($|x+y| \leq |x| + |y|$):

$$\begin{aligned} 2\sum_{i=1}^{p} b_i \times (D_{XY} + D_{YZ}) &= \sum_{i=1}^{p} b_i |d_i^X - d_i^Y| + \sum_{i=1}^{p} b_i |d_i^Y - d_i^Z| \\ &= \sum_{i=1}^{p} b_i \left[|d_i^X - d_i^Y| + |d_i^Y - d_i^Z| \right] \\ &\geq \sum_{i=1}^{p} b_i \left(|d_i^X - d_i^Y + d_i^Y - d_i^Z| \right) \\ &= 2\sum_{i=1}^{p} b_i \times D_{XZ} \end{aligned}$$

that is, $D_{XY} + D_{YZ} \geq D_{XZ}$.

B.2 Generalized LUniFrac is Not a Distance

While the Generalized LUniFrac dissimilarity trivially satisfies nonnegativity, identity of indistinguishable elements, and symmetry, it does not satisfy the triangle inequality. We demonstrate this by counterexample. Consider a phylogenetic tree with three taxa, as displayed in Figure B.1. Suppose the relative abundance of these three taxa has been observed for three subjects at two time points each, as shown in Table B.1. The resulting weighted LUniFrac dissimilarity matrix is shown in Table B.2. From the table, $D_{12} + D_{23} = 0.108 + 0.132 = 0.239 < 0.278 = D_{13}$. That is, this dissimilarity does not

satisfy the triangle inequality because it does not hold that $D_{AB} + D_{BC} \geq D_{AC}$ for every combination of subjects A, B, C. Therefore, the weighted LUniFrac dissimilarity is not a proper distance.

Figure B.1: Simple phylogenetic tree with three taxa. Numbers indicate branch lengths.

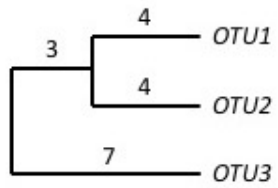


Table B.1: Relative abundances of taxa in counterexample.

Subject	Time	OTU1	OTU2	OTU3
Subj1	1	0.40	0.10	0.50
Subj1	2	0.05	0.65	0.30
Subj2	1	0.30	0.05	0.65
Subj2	2	0.25	0.25	0.5
Subj3	1	0.55	0.2	0.25
Subj3	2	0.45	0.15	0.4

Table B.2: Weighted LUniFrac dissimilarity for taxon abundances in Table B.1.

	Subj1	Subj2	Subj3
Subj1	0	0.108	0.278
Subj2	0.108	0	0.132
Subj3	0.278	0.132	0

Appendix C

APPENDIX TO CHAPTER 4

C.1 ADMM Algorithm for CSGL

Here, we provide the details of the ADMM algorithm for sparse group compositional lasso. We update each vector of coefficients by minimizing the augmented Lagrangian

$$\mathcal{L}_\mu(\alpha,\beta,\xi)=\frac{1}{2n}\|y-Z\beta\|_2^2+\lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}\|\beta^{(j)}\|_2+\lambda\theta\|\beta\|_1+g_{\mathcal{C}}(\alpha)+\frac{\mu}{2}\|\beta-\alpha+\xi\|_2^2-\frac{\mu}{2}\|\xi\|_2^2$$

with respect to each coefficient vector in turn, where $\beta^{(j)}$ refers to the elements of β in group j; μ is the step size for dual variable updates; and ξ is the scaled Lagrange multiplier. The constraint is encoded as an indicator function $g_{\mathcal{C}}(\beta)$ such that $g_{\mathcal{C}}(\beta)=0$ if $\beta\in\mathcal{C}$ and $g_{\mathcal{C}}(\beta)=\infty$ otherwise, where $\mathcal{C}=\{\beta\in\mathbb{R}^p:\sum_{j=1}^{p}\beta_j=0\}$. In the following, the superscript k is an iteration counter.

To update β, we minimize

$$\mathcal{L}_\mu(\beta;\alpha,\xi)=\frac{1}{2n}\|y-Z\beta\|_2^2+\lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}\|\beta^{(j)}\|_2+\lambda\theta\|\beta\|_1+\frac{\mu}{2}\|\beta-\alpha+\xi\|_2^2 \quad \text{(C.1)}$$

By defining $\tilde{y}=\begin{bmatrix} y \\ \sqrt{n\mu}(\xi-\alpha)\end{bmatrix}\in\mathbb{R}^{(n+p)}$ and $\tilde{Z}=\begin{bmatrix} Z \\ -\sqrt{n\mu}\,\mathbf{I}_p\end{bmatrix}\in\mathbb{R}^{(n+p)\times p}$, (C.1) may be rewritten

$$\mathcal{L}_\mu(\beta;\alpha,\xi)=\frac{1}{2n}\|\tilde{y}-\tilde{Z}\beta\|_2^2+\lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}\|\beta^{(j)}\|_2+\lambda\theta\|\beta\|_1 \quad \text{(C.2)}$$

This is exactly the form of the standard sparse group lasso. Therefore, we use the R package SGL with the augmented data $(\tilde{y},\tilde{Z})$ to perform β updates [114]. This package employs an accelerated generalized gradient descent algorithm.

To update α, we minimize $\mathcal{L}(\alpha;\beta^{k+1},\xi^k)=g_{\mathcal{C}}(\alpha)+\frac{\mu}{2}\|\beta^{k+1}-\alpha+\xi^k\|_2^2$. The solution to this subproblem is the Euclidean projection of $(\beta^{k+1}+\xi^k)$ onto the affine space $\mathcal{C}$.

We update ξ as described in [12] via $\xi^{k+1} = \xi^k + \beta^{k+1} - \alpha^{k+1}$.

C.2 Residual, objective, and dual variable convergence

We demonstrate that the ADMM algorithm used to solve the compositional sparse group lasso problem satisfies the assumptions of Section 3.2.1 in [12]. Under these assumptions, their result states that the ADMM iterates satisfy the following:

- *Residual convergence.* $\beta^k - \alpha^k \to 0$ as $k \to \infty$, i.e., the iterates approach feasibility.
- *Objective convergence.* The objective function $\frac{1}{2}\|y - Z\beta^k\|_2^2 + \lambda(1-\theta)\sum_{j=1}^J \sqrt{p_j}\|\beta^{(j),k}\|_2 + \lambda\theta\|\beta^k\|_1 \to p^\star$ as $k \to \infty$, where $p^\star$ is the optimal value for the problem.
- *Dual variable convergence.* $\xi^k \to \xi^\star$ as $k \to \infty$, where $\xi^\star$ is a dual optimal point.

Proof that assumptions hold: Let $f(\beta) = \frac{1}{2}\|y - Z\beta\|_2^2 + \lambda(1-\theta)\sum_{j=1}^q \sqrt{p_j}\|\beta^{(j)}\|_2 + \lambda\theta\|\beta\|_1$ and $g(\alpha) = g_{\mathcal{C}}(\alpha)$ as previously defined. Let γ be the (unscaled) Lagrange multiplier. The assumptions under which the result holds are as follows:

1. $f(\beta)$ and $g(\alpha)$ are closed, proper, and convex (i.e., there exist β and α, not necessarily unique, to minimize the augmented Lagrangian; note that f and g may be nondifferentiable and may take value $+\infty$), and
2. The unaugmented Lagrangian $\mathcal{L}_0(\alpha, \beta, \xi) = f(\beta) + g(\alpha) + \gamma^\top(\beta - \alpha)$ has a saddle point.

To verify Assumption 1, first note that $f(\beta)$ is a linear combinations of norms, so it is convex by the triangle inequality and closed because it is continuous on a closed domain ($\mathbb{R}^p$). To show that $f(\beta)$ is proper, note that it is everywhere $> -\infty$: since $x^2 \geq 0$ everywhere and $|x| \geq 0$ everywhere, their sums and square roots are also ≥ 0 everywhere, so therefore $f(\beta)$ is strictly $> -\infty$ everywhere. Also, $f(\beta)$ is somewhere $< \infty$. When $\beta = 0$, $\frac{1}{2}\|y\|_2^2 + \|0\|_2 + \|0\|_1 =$

$\frac{1}{2}\|y\|_2^2 + 0$, which is finite because every element of y is finite. The characteristic function $g_{\mathcal{C}}(\alpha)$ is a proper convex function, since $g_{\mathcal{C}}(\alpha) = 0$ if $\alpha \in \mathcal{C}$ and $+\infty$ otherwise. It is closed because $\mathcal{C}$ is a closed set. Hence Assumption 1 holds.

To verify Assumption 2, the (unscaled) unaugmented Lagrangian is

$$\mathcal{L}_0(\alpha, \beta, \gamma) = \frac{1}{2}\|y - Z\beta\|_2^2 + \lambda(1-\theta)\sum_{j=1}^{q}\sqrt{p_j}\|\beta^{(j)}\|_2 + \lambda\theta\|\beta\|_1 + g_{\mathcal{C}}(\alpha) + \gamma^\top(\beta - \alpha).$$

(For the scaled version of the augmented Lagrangian in the main text, $\xi = \gamma/\mu$.) We must show that this has a saddle point. In this case, the optimization problem is the Lagrange function of a solvable convex problem and the Slater condition is satisfied. Specifically, $\mathcal{L}_0$ is convex in α, β for every fixed $\gamma \in \mathbb{R}_+^{(p+1)}$ and linear in γ for every fixed (α, β) (therefore concave in γ). Then $\mathcal{L}_0$ has a saddle point that coincides with the optimal values of (α, β) and γ (by, e.g., Section VII.4 of [**?**]).

Therefore the assumptions are verified and the ADMM algorithm is guaranteed residual, objective, and dual convergence.

C.3 Sign Consistency and Bounded Loss

We demonstrate that with high probability there exists an optimal solution $\hat{\beta}$ to the CSGL optimization problem that satisfies sign consistency and bounded loss. We require fairly standard assumptions on the signal strength and the correlation between columns of the design matrix.

For the purposes of this proof, we use the non-symmetric form of the problem. We assume without loss of generality that the last component, Z_p, is chosen as the reference, that it belongs to the last group (q), and that $\hat{\beta}_p \neq 0$. As in the main text, Z^p indicates the matrix of log-ratios excluding the reference component, so that $Z^p_{ij} = Z_{ij}/Z_{ip}$ for $j \in \{1, \ldots, p-1\}$. $\beta_{\setminus p}$ is the $(p-1)$-vector of coefficients for the non-symmetric linear log contrast model, with estimates $\hat{\beta}_{\setminus p}$. The full coefficient vector, used in the penalty terms, is $D\beta_{\setminus p} = \left(\beta_{\setminus p}, -1_{p-1}^\top \beta_{\setminus p}\right) = \beta$.

The non-symmetric form of the CSGL problem is

$$\hat{\beta}_{\setminus p} = \underset{\beta_{\setminus p}}{\operatorname{argmin}} \left(\frac{1}{2n} \|y - Z^p \beta_{\setminus p}\|_2^2 + \lambda_1 \sum_{j=1}^{q} \sqrt{p_j} \|(D\beta_{\setminus p})^{(j)}\|_2 + \lambda_2 \|D\beta_{\setminus p}\|_1 \right) \tag{C.3}$$

where $D = \begin{bmatrix} \mathbf{I}_{p-1} & -1_{p-1} \end{bmatrix}^\top \in \mathbb{R}^{(p \times p-1)}$ to enforce the constraint $\beta_p = -\sum_{k=1}^{p} \beta_k$, and $(D\beta_{\setminus p})^{(j)}$ indicates the elements of the vector $(\beta_1, \ldots, \beta_{p-1}, \beta_p = -1_{p-1}^\top \beta_{\setminus p})$ corresponding to group j. For notational simplicity we denote $\lambda_1 = \lambda(1-\theta)$ and $\lambda_2 = \lambda\theta$.

Notation

As in the main text, the following sets will be useful. Let $\beta_k^{(j)}$ denote the kth feature in the jth group.

- $\mathcal{G} = \{1 \leq j \leq q : \beta^{(j)} \neq 0\}$ is the set of groups with at least one truly nonzero element;
- $\mathcal{S}_j = \{1 \leq k \leq p_j : \beta_k^{(j)} \neq 0\}$ indicates nonzero features in group j; and
- $\mathcal{S}_\mathcal{G} = \bigcup_{j \in \mathcal{G}} \mathcal{S}_j$ indicates nonzero features in any group.

The magnitude of each set is indicated by $|\mathcal{G}| = g$, $|\mathcal{S}_j| = s_j$, and $|\mathcal{S}_\mathcal{G}| = s$. The equivalent sets for $\hat{\beta}$ are indicated by $\hat{\mathcal{G}}$, $\hat{\mathcal{S}}_j$, and $\hat{\mathcal{S}}_{\hat{\mathcal{G}}}$. Complements are indicated by a superscript c; for example, $\mathcal{G}^c$ refers to groups that are uniformly zero. A superscript p indicates that the reference element is excluded from the set. For example, $\hat{\mathcal{S}}_{\hat{\mathcal{G}}}^p$ is the set of indices for all nonzero $\hat{\beta}_{\setminus p}$.

The three cases for zero or nonzero coefficients can be described using these three sets. Let $\beta_k^{(j)}$ indicate the coefficient for feature k in group j.

- *Case 1: Nonzero betas in nonzero groups.* When $\beta_k^{(j)} \neq 0$, $j \in \mathcal{G}$ and $k \in \mathcal{S}_j$, and the element is included in $\mathcal{S}_\mathcal{G}$. This case includes s coefficients (or $s-1$ when the reference is excluded, i.e., elements in $\mathcal{S}_\mathcal{G}^p$).

- *Case 2: Zero betas in nonzero groups.* When $\beta_k^{(j)} = 0$ but other elements of the group are nonzero (so that $\beta^{(j)} \neq 0$), then $j \in \mathcal{G}$ and $k \in \mathcal{S}_j^c$. This set is indicated by $\mathcal{S}_{\mathcal{G}}^c = \bigcup_{j \in \mathcal{G}} \mathcal{S}_j^c$. The magnitude of this set is $r = |\mathcal{S}_{\mathcal{G}}^c|$.

- *Case 3: Uniformly zero groups.* For groups such that $\beta^{(j)} = 0$ for all elements, $j \in \mathcal{G}^c$. The magnitude of this set is $t = |\mathcal{S}_{\mathcal{G}^c}^c| = \sum_{j \in \mathcal{G}^c} p_j$. Note that since there are no nonzero elements in uniformly zero groups, $\mathcal{S}_{\mathcal{G}^c} = \varnothing$.

Every element of β falls into one of these three categories, so $s + r + t = p$.

Subgradient conditions (Lemma 3.1).

The quantity $\hat{\beta}$ is a solution of the CSGL problem if and only if the following subgradient conditions are satisfied for $(\hat{\beta}_1, \ldots, \hat{\beta}_{p-1})$ and $\hat{\beta}_p = -\mathbf{1}_{p-1}^\top \hat{\beta}_{\setminus p}$. Let $\hat{B}_{\hat{\mathcal{S}}_{\hat{\mathcal{G}}}}$ be the block-diagonal matrix with g blocks defined by $\left\{ \mathbf{I}_{\hat{s}_j} \frac{\sqrt{p_j}}{\|\hat{\beta}^{(j)}\|_2}, j \in \hat{\mathcal{G}} \right\}$ and $\hat{b}_q = \frac{\sqrt{p_q}}{\|\hat{\beta}^{(q)}\|_2}$. Then the subgradient conditions are:

$$-\frac{1}{n}(Z^p_{\hat{\mathcal{S}}^p_{\hat{\mathcal{G}}}})^\top (y - Z^p \hat{\beta}_{\setminus p}) + \lambda_1 \left\{ \hat{B}_{\hat{\mathcal{S}}^p_{\hat{\mathcal{G}}}} \hat{\beta}_{\hat{\mathcal{S}}^p_{\hat{\mathcal{G}}}} - \hat{b}_q \hat{\beta}_p \mathbf{1}_{s-1} \right\} + \lambda_2 \left\{ \operatorname{sgn}(\hat{\beta}_{\hat{\mathcal{S}}^p_{\hat{\mathcal{G}}}}) - \operatorname{sgn}(\hat{\beta}_p) \mathbf{1}_{s-1} \right\} = 0 \tag{A1}$$

$$\max_{j \in \hat{\mathcal{G}}} \left(\left\| \frac{1}{n} (Z^{p,(j)}_{\hat{\mathcal{S}}^c_j})^\top (y - Z^p \hat{\beta}_{\setminus p}) + (\lambda_1 \hat{b}_q \hat{\beta}_p + \lambda_2 \operatorname{sgn}(\hat{\beta}_p)) \mathbf{1}_{(p_j - s_j)} \right\|_\infty \right) \leq \lambda_2 \tag{A2}$$

$$\max_{j \in \hat{\mathcal{G}}^c} \left(\frac{1}{\sqrt{p_j}} \left\| S\left(\frac{1}{n} Z^{(j)\top} (y - Z^p \hat{\beta}_{\setminus p}) + (\lambda_1 \hat{b}_q \hat{\beta}_p + \lambda_2 \operatorname{sgn}(\hat{\beta}_p)) \mathbf{1}_{p_j}, \lambda\theta \right) \right\|_2 \right) \leq \lambda_1 \tag{A3}$$

Note that (A1) corresponds to nonzero features in nonzero groups, i.e., $\hat{\beta}_k^{(j)} \neq 0$ (Case 1); (A2) corresponds to zero features in nonzero groups, i.e., $\hat{\beta}^{(j)} \neq 0$ and $\hat{\beta}_k^{(j)} = 0$ (Case 2); and (A3) corresponds to groups that are uniformly zero, i.e., $\hat{\beta}^{(j)} = 0$ (Case 3).

Proof: All terms of the criterion are convex, so we can use subgradient calculations to derive

the optimality criteria. Specifically, $x \in \mathbb{R}^p$ is a minimum of f if and only if $0 \in \partial f|_x$. The subgradient is equal to the gradient if the function is differentiable at x.

The first term of (C.3) is continuously differentiable with derivative

$$\frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\frac{1}{2n} \|y - Z^p \hat{\beta}_{\setminus p}\|_2^2 \right) = -\frac{1}{n} (Z_k^{p,(j)})^\top (y - Z^p \hat{\beta}_{\setminus p}),$$

for each element $\hat{\beta}_k^{(j)}$ of $\hat{\beta}_{\setminus p}$, and this term is included in each of (A1)-(A3). We now consider the second and third terms, which distinguish the three cases.

In Case 1, $\hat{\beta}_k^{(j)} \neq 0$, which also implies that $\hat{\beta}^{(j)}$ is not identically zero. In this case the second term of (C.3) is differentiable with derivative given by

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\lambda_1 \sum_{l=1}^{q} \sqrt{p_l} \|\hat{\beta}^{(l)}\|_2 \right) &= \lambda_1 \frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left[\sqrt{p_j} \|\hat{\beta}^{(j)}\|_2 + \sqrt{p_q} \left\{ \sum_{k=1}^{p_q - 1} (\hat{\beta}_k^{(q)})^2 + (-1_{p-1}^\top \hat{\beta}_{\setminus p})^2 \right\}^{1/2} \right] \\
&= \lambda_1 \left\{ \frac{\sqrt{p_j} \hat{\beta}_k^{(j)}}{\|\hat{\beta}^{(j)}\|_2} + \frac{\sqrt{p_q}}{2\|\hat{\beta}^{(q)}\|_2} \cdot 2(-1_{p-1}^\top \hat{\beta}_{\setminus p})(-1) \right\} \\
&= \lambda_1 \left(\frac{\sqrt{p_j} \hat{\beta}_k^{(j)}}{\|\hat{\beta}^{(j)}\|_2} - \frac{\sqrt{p_q} \hat{\beta}_p}{\|\hat{\beta}^{(q)}\|_2} \right)
\end{aligned}$$

for all $\hat{\beta}_k^{(j)}$ in $\hat{\beta}_{\setminus p}$. The third term is also differentiable, with derivative

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\lambda_2 \|\hat{\beta}\|_1 \right) &= \lambda_2 \frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(|\hat{\beta}_k^{(j)}| + |\hat{\beta}_p| \right) \\
&= \lambda_2 \frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left\{ |\hat{\beta}_k^{(j)}| + \left| (-1_{p-1}^\top \hat{\beta}_{\setminus p}) \right| \right\} \\
&= \lambda_2 \left\{ \operatorname{sgn}(\hat{\beta}_k^{(j)}) - \operatorname{sgn}(\hat{\beta}_p) \right\}.
\end{aligned}$$

Combining these provides the Case 1 result.

On the other hand, if $\hat{\beta}_k^{(j)} = 0$, but other elements of the group $\hat{\beta}^{(j)}$ are nonzero, we have for the third term

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\lambda_2 \|\hat{\beta}\|_1 \right) &= \lambda_2 \frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(|\hat{\beta}_k^{(j)}| + |\hat{\beta}_p| \right) \\
&= \lambda_2 \left\{ u - \operatorname{sgn}(\hat{\beta}_p) \right\}
\end{aligned}$$

where $u \in \{u : |u| \leq 1\}$. With a little bit of algebra, this provides the Case 2 result.

Finally, if $\hat{\beta}^{(j)} = 0$, then we have for the second term

$$\begin{aligned}\frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\lambda_1 \sum_{l=1}^{q} \sqrt{p_l} \|\hat{\beta}^{(l)}\|_2 \right) &= \lambda_1 \frac{\partial}{\partial \hat{\beta}_k^{(j)}} \left(\sqrt{p_k} \|\hat{\beta}^{(k)}\|_2 + \sqrt{p_q} \|\hat{\beta}^{(q)}\|_2 \right) \\ &= \lambda_1 \left(\sqrt{p_k} v - \frac{\sqrt{p_q} \hat{\beta}_p}{\|\hat{\beta}^{(q)}\|_2} \right)\end{aligned}$$

where $v \in \{v : \|v\|_2 \leq 1\}$. The third term here is the same as in Case 2, providing the Case 3 result.

Probability Statements

Let $\Sigma^p = \frac{1}{n}(Z^p)^\top Z^p$ be the sample covariance matrix for the log-ratio transformed data with reference element p. Submatrices are indicated by subscripts. Thus, for example, $\Sigma^p_{\mathcal{S}_\mathcal{G}^p \mathcal{S}_\mathcal{G}^p}$ indicates the submatrix of the sample covariance including entries (i, j) such that $i \in \mathcal{S}_\mathcal{G}^p, j \in \mathcal{S}_\mathcal{G}^p$. With probability at least $1 - 2p \exp\{-n \min(\theta, 1-\theta)^2 \eta^2 \lambda^2 / (8\kappa\sigma^2)\}$, the following statements hold:

$$\left\| n^{-1} Z^\top_{\mathcal{S}_\mathcal{G}} \epsilon \right\|_\infty \leq \frac{\lambda}{2} = \frac{\lambda_1 + \lambda_2}{2} \tag{P1}$$

$$\left\| n^{-1} (Z^p_{\mathcal{S}^c})^\top \epsilon - n^{-1} \Sigma^p_{\mathcal{S}^c \mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p \mathcal{S}_\mathcal{G}^p})^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^p})^\top \epsilon \right\|_\infty \leq \frac{\lambda \min(\theta, 1-\theta)\eta}{\sqrt{\kappa}} = \frac{\min(\lambda_1, \lambda_2)\,\eta}{\sqrt{\kappa}} \tag{P2}$$

Proof: Assume that $\epsilon \sim N(0, \sigma^2)$ for all subjects and that the columns of Z have been standardized such that $\max_j \|Z_j\|_2^2 \leq n$.

Proof of (P1): Using Boole's inequality and Gaussian tail bounds,

$$\begin{aligned}P\left(\left\| n^{-1} Z^\top_{\mathcal{S}_\mathcal{G}} \epsilon \right\|_\infty \geq \lambda/2 \right) &\leq \sum_{j \in \mathcal{S}_\mathcal{G}} P\left(|n^{-1} Z_j^\top \epsilon| \geq \lambda/2 \right) \\ &\leq \sum_{j \in \mathcal{S}_\mathcal{G}} 2 \exp\left\{ -\frac{\lambda^2/4}{2\sigma^2/n} \right\} \\ &= 2s \exp\{-n\lambda^2/(8\sigma^2)\}\end{aligned}$$

where we have used the fact that $\frac{1}{n}Z_j^\top \epsilon$ is distributed Normally with mean 0 and variance no larger than $\frac{\sigma^2}{n}$ for each column Z_j of $Z_{\mathcal{S}}$ based upon our chosen standardization. Therefore $P(\|n^{-1}Z_{\mathcal{S}_\mathcal{G}}^\top \epsilon\|_\infty \leq \lambda/2) \geq 1 - 2s\exp\{-n\lambda^2/(8\sigma^2)\}$. This is identical to the first inequality of (A3) in [74].

Proof of (P2): Let $\mathbf{H}$ be the hat (projection) matrix $Z^p_{\mathcal{S}^p_\mathcal{G}}\{(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top Z^p_{\mathcal{S}^p_\mathcal{G}}\}^{-1}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top$. Then $(\mathbf{I} - \frac{1}{n}\mathbf{H})$ is also a projection matrix and therefore has maximum eigenvalue 1. Using this, along with Boole's inequality and Gaussian tail bounds:

$$\begin{aligned}
P\Bigg(\left\|n^{-1}(Z^p_{\mathcal{S}^c})^\top \epsilon - n^{-1}\Sigma_{\mathcal{S}^c\mathcal{S}^p_\mathcal{G}}\Sigma^{-1}_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top\epsilon\right\|_\infty \geq \frac{\min(\lambda_1,\lambda_2)\,\eta}{\sqrt{\kappa}}\Bigg) \\
&= P\left(\left\|n^{-1}(Z^p_{\mathcal{S}^c_\mathcal{G}})^\top(\mathbf{I} - n^{-1}\mathbf{H})\epsilon\right\|_\infty \geq \frac{\min(\lambda_1,\lambda_2)\,\eta}{\sqrt{\kappa}}\right) \\
&\leq \sum_{j\in\mathcal{S}^c} P\left(|n^{-1}(Z_j - Z_p)^\top(\mathbf{I} - n^{-1}\mathbf{H})\epsilon| \geq \frac{\min(\lambda_1,\lambda_2)\,\eta}{\sqrt{\kappa}}\right) \\
&\leq 2(r+t)\exp\left\{-\frac{(\min(\lambda_1,\lambda_2)\,\eta)^2}{8\kappa\sigma^2/n}\right\} \\
&= 2(r+t)\exp\{-n\min(\lambda_1,\lambda_2)^2\eta^2/(8\kappa\sigma^2)\}
\end{aligned}$$

since $n^{-1}(Z_j - Z_p)\epsilon$ has maximum variance $4\sigma^2/n$. Therefore, we have that $P\Big\{\left\|n^{-1}(Z^p_{\mathcal{S}^c_\mathcal{G}})^\top\epsilon - n^{-1}\Sigma_{\mathcal{S}^c_\mathcal{G}\mathcal{S}^p_\mathcal{G}}\Sigma^{-1}_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top\epsilon\right\|_\infty \leq \lambda_2\eta\Big\} \geq 1 - 2(r+t)\exp\{-n\min(\lambda_1,\lambda_2)^2\eta^2/(8\kappa\sigma^2)\}$. This corresponds to the second inequality of (A3) in [74].

Combined Probability of (P1) and (P2): Return to the original penalty parameters, $\lambda_1 = \lambda(1-\theta)$ and $\lambda_2 = \lambda\theta$ for some predetermined value of θ. Then

$$\begin{aligned}
P(\text{P1 and P2}) &\geq 1 - \left[2s\exp\{-n\lambda^2/(8\sigma^2)\}\right] - \left[2(r+t)\exp\{-n\lambda^2\min(\theta,1-\theta)^2\eta^2/(8\kappa\sigma^2)\}\right] \\
&\geq 1 - 2(s+r+t)\exp\{-n\lambda^2\min(\theta,1-\theta)^2\eta^2/(8\kappa\sigma^2)\}
\end{aligned}$$

That is, the combined probability of (P1) and (P2) can be written as $1-2p\,\exp\{-n\min(\theta,1-\theta)^2\eta^2\lambda^2/(8\kappa\sigma^2)\}$. Condition (C3) on the minimum value of λ ensures that this combined probability is non-negative.

Required Conditions

Condition 1: Irrepresentable Conditions (Assumption 3.2 in main text): These conditions are in the spirit of the Irrepresentable Condition for the lasso and group lasso. We assume that θ has been chosen in advance and is fixed, with $\lambda_1 = \lambda(1-\theta)$ and $\lambda_2 = \lambda\theta$, and $\kappa = \max_j \sqrt{p_j}$ is the maximum group size. Then there exists some $\eta \in (0, 1]$ such that

$$\sqrt{\kappa}\left\|\Sigma^p_{\mathcal{S}^c\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left\{\operatorname{sgn}(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - \operatorname{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\} + \operatorname{sgn}(\beta_p)\mathbf{1}_{(r+t)}\right\|_\infty \leq \frac{(1-\eta)}{2}\cdot\frac{\min(\theta, 1-\theta)}{\max(\theta, 1-\theta)} \tag{C1a}$$

$$\sqrt{\kappa}\left\|\Sigma^p_{\mathcal{S}^c\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}}\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}\right) + b_q\beta_p\mathbf{1}_{(r+t)}\right\|_\infty \leq \frac{(1-\eta)}{2}\cdot\frac{\min(\theta, 1-\theta)}{\max(\theta, 1-\theta)} \tag{C1b}$$

Here, $\mathcal{S}^c = \mathcal{S}^c_{\mathcal{G}^c} \bigcup \mathcal{S}^c_{\mathcal{G}}$ includes all features with truly zero coefficients, regardless of whether the entire group is zero or nonzero.

Condition 2: Minimum Signal Size: The minimum signal size is

$$\beta_{\min} > \frac{\psi}{2}\left\{(2\kappa+1)\lambda_1 + 3\lambda_2\right\} \tag{C2}$$

where $\psi = \left\|D_{\mathcal{S}_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}D^\top_{\mathcal{S}_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}\right\|_\infty$.

Condition 3: Regularization Parameter: In order for the combined bound in Lemma 2 to hold, λ must satisfy

$$\lambda = \frac{c_1\sigma\sqrt{\kappa}}{\eta\,\min(\theta, 1-\theta)}\left\{\frac{\log(2p)}{n}\right\}^{1/2} \tag{C3}$$

for some constant $c_1 > 2\sqrt{2}$.

Estimation Consistency and Error Bounds (Theorem 3.3)

Assume that Condition 1 holds, the minimum signal size satisfies Condition 2, and the regularization parameter λ satisfies Condition 3 for the selected θ. Then with probability at least $1 - 2p\exp\{-n\min(\theta, 1-\theta)^2\eta^2\lambda^2/(8\kappa\sigma^2)\}$, problem (C.3) has an optimal solution $\hat{\beta}$ that has the following two properties:

1. Sign consistency: $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta)$, and

2. Bounded ℓ_∞ loss: $\|\hat{\beta}_{\mathcal{S}_\mathcal{G}} - \beta_{\mathcal{S}_\mathcal{G}}\|_\infty \leq \frac{\psi}{2}\{(2\kappa+1)\lambda_1 + 3\lambda_2\}$.

Proof:

The broad outline of this proof is that, conditional on the events defined in Lemma 2, we will use deterministic reasoning to find some $\hat{\beta}$ with the desired properties such that (A1)-(A3) hold. This proof is adapted from Lin et al. (2014) [74]. If $\lambda_2 = 0$, the CSGL problem reduces to the compositional group lasso and (A2) does not apply (there are no zero features in nonzero groups). Conversely, if $\lambda_1 = 0$, the CSGL problem reduces to the compositional lasso and the proof in [74] applies exactly. We therefore assume $\lambda_2 \neq 0$ and $\lambda_1 \neq 0$.

Condition on (P1) and (P2) holding, and take $\hat{\beta}_{\mathcal{S}_\mathcal{G}^c} = 0$. We first consider Case 1 (nonzero $\hat{\beta}_k^{(j)}$, for which $j \in \mathcal{G}$ and $k \in \mathcal{S}_j$) and subgradient condition (A1). For each group $j \in \mathcal{G}$, substituting $y = Z^p_{\mathcal{S}_j}\beta_{\setminus p} + \epsilon$ and replacing $\hat{\mathcal{S}}^p_\mathcal{G}$ with $\mathcal{S}^p_\mathcal{G}$ in (A1) yields

$$
\begin{aligned}
0 &= -\frac{1}{n}(Z^p_{\mathcal{S}_j^p})^\top(Z^p\beta_{\setminus p} + \epsilon - Z^p\hat{\beta}_{\setminus p}) + \lambda_1\left(\hat{B}_{\mathcal{S}_j^p}\hat{\beta}^{(j)}_{\mathcal{S}_j^p} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s_j}\right) + \lambda_2\left\{\text{sgn}(\hat{\beta}_{\mathcal{S}_j^p}) - \text{sgn}(\hat{\beta}_p)\mathbf{1}_{s_j}\right\} \\
&\iff \frac{1}{n}(Z^p_{\mathcal{S}_j^p})^\top Z^p_{\mathcal{S}^p_\mathcal{G}}(\beta_{\mathcal{S}^p_\mathcal{G}} - \hat{\beta}_{\mathcal{S}^p_\mathcal{G}}) + \frac{1}{n}(Z^p_{\mathcal{S}_j^p})^\top\epsilon = \lambda_1\left(\hat{B}_{\mathcal{S}_j^p}\hat{\beta}^{(j)}_{\mathcal{S}_j^p} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s_j}\right) + \lambda_2\left\{\text{sgn}(\hat{\beta}_{\mathcal{S}_j^p}) - \text{sgn}(\hat{\beta}_p)\mathbf{1}_{s_j}\right\} \\
&\iff \Sigma^p_{\mathcal{S}_j^p\mathcal{S}^p_\mathcal{G}}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}} - \beta_{\mathcal{S}^p_\mathcal{G}}) = \frac{1}{n}(Z^p_{\mathcal{S}_j^p})^\top\epsilon - \lambda_1\left(\hat{B}_{\mathcal{S}_j^p}\hat{\beta}^{(j)}_{\mathcal{S}_j^p} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s_j}\right) - \lambda_2\left\{\text{sgn}(\hat{\beta}_{\mathcal{S}_j^p}) - \text{sgn}(\hat{\beta}_p)\mathbf{1}_{s_j}\right\}
\end{aligned}
$$

so that

$$
\hat{\beta}_{\mathcal{S}^p_\mathcal{G}} - \beta_{\mathcal{S}^p_\mathcal{G}} = (\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}\left[\frac{1}{n}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top\epsilon - \lambda_1\left(\hat{B}_{\mathcal{S}^p_\mathcal{G}}\hat{\beta}_{\mathcal{S}^p_\mathcal{G}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}\right) - \lambda_2\left\{\text{sgn}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}) - \text{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\}\right] \tag{C.4}
$$

Define $\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}$ by Equation (C.4) with $\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}$, $\hat{\beta}_p$, $\text{sgn}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}})$ and $\text{sgn}(\beta_p)$ replaced with their true values $\beta_{\mathcal{S}^p_\mathcal{G}}$, β_p, $\text{sgn}(\beta_{\mathcal{S}^p_\mathcal{G}})$, and $\text{sgn}(\beta_p)$. We will show that this choice of $\hat{\beta}$ satisfies the subgradient criteria and has the desired properties.

Multiplying by $D_{\mathcal{S}_\mathcal{G}\mathcal{S}^p_\mathcal{G}}$ on both sides to transform the nonzero elements of $\beta_{\setminus p}$ to the nonzero elements of β (where the subscripts on D indicate rows in $\mathcal{S}_\mathcal{G}$ and columns in $\mathcal{S}^p_\mathcal{G}$, respectively), taking the ℓ_∞ norm of each side, and applying (P1) and the triangle inequality,

we have

$$\begin{aligned}
\|\hat{\beta}_{\mathcal{S}_\mathcal{G}} - \beta_{\mathcal{S}_\mathcal{G}}\|_\infty &= \Big\| n^{-1} D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} \Big[(Z^p_{\mathcal{S}_\mathcal{G}^p})^\top \epsilon - \lambda_1 \left(\hat{B}_{\mathcal{S}_\mathcal{G}^p} \hat{\beta}_{\mathcal{S}_\mathcal{G}^p} - \hat{b}_q \hat{\beta}_p \mathbf{1}_{s-1} \right) \\
&\qquad - \lambda_2 \left\{ \operatorname{sgn}(\hat{\beta}_{\mathcal{S}_\mathcal{G}^p}) - \operatorname{sgn}(\hat{\beta}_p) \mathbf{1}_{s-1} \right\} \Big] \Big\|_\infty \\
&\leq \left\| n^{-1} D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^p})^\top \epsilon \right\|_\infty + \lambda_1 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} \left(\hat{B}_{\mathcal{S}_\mathcal{G}^p} \hat{\beta}_{\mathcal{S}_\mathcal{G}^p} - \hat{b}_q \hat{\beta}_p \mathbf{1}_{s-1} \right) \right\|_\infty \\
&\qquad + \lambda_2 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} \left\{ \operatorname{sgn}(\hat{\beta}_{\mathcal{S}_\mathcal{G}^p}) - \operatorname{sgn}(\hat{\beta}_p) \mathbf{1}_{s-1} \right\} \right\|_\infty \\
&= \left\| n^{-1} D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} Z^\top_{\mathcal{S}_\mathcal{G}} \epsilon \right\|_\infty + \lambda_1 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \hat{B}_{\mathcal{S}_\mathcal{G}} \hat{\beta}_{\mathcal{S}_\mathcal{G}} \right\|_\infty \\
&\qquad + \lambda_2 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \operatorname{sgn}(\hat{\beta}_{\mathcal{S}_\mathcal{G}}) \right\|_\infty \\
&\leq \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \left\| n^{-1} Z^\top_{\mathcal{S}_\mathcal{G}} \epsilon \right\|_\infty + \lambda_1 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \left\| \hat{B}_{\mathcal{S}_\mathcal{G}} \hat{\beta}_{\mathcal{S}_\mathcal{G}} \right\|_\infty \\
&\qquad + \lambda_2 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \left\| \operatorname{sgn}(\hat{\beta}_{\mathcal{S}_\mathcal{G}}) \right\|_\infty \\
&\leq \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \left\| n^{-1} Z^\top_{\mathcal{S}_\mathcal{G}} \epsilon \right\|_\infty + \lambda_1 \kappa \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \\
&\qquad + \lambda_2 \left\| D_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} D^\top_{\mathcal{S}_\mathcal{G}\mathcal{S}_\mathcal{G}^p} \right\|_\infty \\
&\leq \frac{\psi}{2} \{ (2\kappa + 1)\lambda_1 + 3\lambda_2 \} < \beta_{\min}
\end{aligned}$$

by Condition 2, where we use that $|\hat{\beta}_k^{(j)}| / \|\hat{\beta}_k^{(j)}\|_2 \leq 1$ and, as before, $\kappa = \max_j \sqrt{p_j}$. Therefore this $\hat{\beta}$ satisfies (A1) and has the desired properties of sign consistency and bounded ℓ_∞ loss. However, we must still verify that $\hat{\beta}$ satisfies (A2) and (A3).

For (A2), we are in the setting where $\hat{\beta}_k^{(j)} = 0$ within a nonzero group $\hat{\beta}^{(j)}$. This corresponds to features in the set $\mathcal{S}_\mathcal{G}^c = \{j, k : j \in \mathcal{G}, k \in \mathcal{S}_j^c\}$). Using $(C.4)$ and replacing y by $Z^p_{\mathcal{S}_\mathcal{G}^p} \beta_{\mathcal{S}_\mathcal{G}^p} + \epsilon$, the interior of $(A2)$ becomes

$$\begin{aligned}
W &\equiv n^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^c})^\top (y - Z^p \hat{\beta}_{\mathcal{S}_\mathcal{G}^p}) + \{ \lambda_1 b_q \beta_p + \lambda_2 \operatorname{sgn}(\beta_p) \} \mathbf{1}_r \\
&= n^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^c})^\top (Z^p_{\mathcal{S}_\mathcal{G}^p} \beta_{\mathcal{S}_\mathcal{G}^p} + \epsilon - Z^p \hat{\beta}_{\mathcal{S}_\mathcal{G}^p}) + \{ \lambda_1 b_q \beta_p + \lambda_2 \operatorname{sgn}(\beta_p) \} \mathbf{1}_r \\
&= n^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^c})^\top Z^p_{\mathcal{S}_\mathcal{G}^p} (\beta_{\mathcal{S}_\mathcal{G}^p} - \hat{\beta}_{\mathcal{S}_\mathcal{G}^p}) + n^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^c})^\top \epsilon + \{ \lambda_1 b_q \beta_p + \lambda_2 \operatorname{sgn}(\beta_p) \} \mathbf{1}_r .
\end{aligned}$$

Substituting Equation (C.4) for $(\beta_{\mathcal{S}_\mathcal{G}^p} - \hat{\beta}_{\mathcal{S}_\mathcal{G}^p})$,

$$W = -n^{-1} (Z^p_{\mathcal{S}_\mathcal{G}^c})^\top Z^p_{\mathcal{S}_\mathcal{G}^p} (\Sigma^p_{\mathcal{S}_\mathcal{G}^p\mathcal{S}_\mathcal{G}^p})^{-1} \left\{ \frac{1}{n} (Z^p_{\mathcal{S}_\mathcal{G}^p})^\top \epsilon \right\}$$

$$
\begin{aligned}
&+ n^{-1}(Z^p_{\mathcal{S}^c_{\mathcal{G}}})^\top Z^p_{\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left\{\lambda_1\left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}}\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}\right)\right\}\\
&+ n^{-1}(Z^p_{\mathcal{S}^c_{\mathcal{G}}})^\top Z^p_{\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left[\lambda_2\left\{\operatorname{sgn}(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - \operatorname{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\}\right]\\
&+ n^{-1}(Z^p_{\mathcal{S}^c_{\mathcal{G}}})^\top\epsilon + \{\lambda_1 b_q\beta_p + \lambda_2\operatorname{sgn}(\beta_p)\}\mathbf{1}_r\\
= &-n^{-1}\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}(Z^p_{\mathcal{S}^p_{\mathcal{G}}})^\top\epsilon + \lambda_1\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}}\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}\right)\\
&+ \lambda_2\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left\{\operatorname{sgn}(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - \operatorname{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\}\\
&+ n^{-1}(Z^p_{\mathcal{S}^c_{\mathcal{G}}})^\top\epsilon + \{\lambda_1 b_q\beta_p + \lambda_2\operatorname{sgn}(\beta_p)\}\mathbf{1}_r
\end{aligned}
$$

Then, taking the ℓ_∞ norm of both sides and using (P2), Conditions (C1a) and (C1b), and the triangle inequality, we have

$$
\begin{aligned}
\|W\|_\infty &\le \left\|n^{-1}(Z^p_{\mathcal{S}^c_{\mathcal{G}}})^\top\epsilon - n^{-1}\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}(Z^p_{\mathcal{S}^p_{\mathcal{G}}})^\top\epsilon\right\|_\infty\\
&\quad + \lambda_2\left\|\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left\{\operatorname{sgn}(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - \operatorname{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\} + \operatorname{sgn}(\beta_p)\mathbf{1}_r\right\|_\infty\\
&\quad + \lambda_1\left\|\Sigma^p_{\mathcal{S}^c_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}}\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}\right) + b_q\beta_p\mathbf{1}_r\right\|_\infty\\
&\le \frac{\min(\lambda_1,\lambda_2)\eta}{\sqrt{\kappa}} + \frac{\lambda_2(1-\eta)}{2\sqrt{\kappa}}\left\{\frac{\min(\lambda_1,\lambda_2)}{\max(\lambda_1,\lambda_2)}\right\} + \frac{\lambda_1(1-\eta)}{2\sqrt{\kappa}}\left\{\frac{\min(\lambda_1,\lambda_2)}{\max(\lambda_1,\lambda_2)}\right\}\\
&\le \frac{\lambda_2\eta}{\sqrt{\kappa}} + \frac{\lambda_2(1-\eta)}{\sqrt{\kappa}} \le \lambda_2
\end{aligned}
$$

because $\mathcal{S}^c_{\mathcal{G}} \subset \mathcal{S}^c$. This verifies (A2).

For (A3), we are in the setting where an entire group $\hat{\beta}^{(j)}$ is zero. For $j \in \mathcal{G}^c$ and using Equation (C.4), we define V_j via

$$
\begin{aligned}
\sqrt{p_j}V_j &\equiv \left\|S\left((Z^p_{\mathcal{S}^c_j})^\top(y - Z^p\hat{\beta}_{\setminus p})/n + \lambda_2\operatorname{sgn}(\beta_p)\mathbf{1}_{p_j} + \lambda_1 b_q\beta_q\mathbf{1}_{p_j}, \lambda\theta\right)\right\|_2\\
&\le \left\|n^{-1}(Z^p_{\mathcal{S}^c_j})^\top(Z^p_{\mathcal{S}^p_{\mathcal{G}}}\beta_{\mathcal{S}^p_{\mathcal{G}}} + \epsilon - Z^p\hat{\beta}_{\setminus p}) + \lambda_2\operatorname{sgn}(\beta_p)\mathbf{1}_{p_j} + \lambda_1 b_q\beta_p\mathbf{1}_{p_j}\right\|_2\\
&= \left\|\Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_{\mathcal{G}}}(\beta_{\mathcal{S}^p_{\mathcal{G}}} - \hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) + n^{-1}(Z^p_{\mathcal{S}^c_j})^\top\epsilon + \lambda_2\operatorname{sgn}(\beta_p)\mathbf{1}_{p_j} + \lambda_1 b_q\beta_p\mathbf{1}_{p_j}\right\|_2\\
&= \left\|\Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_{\mathcal{G}}}(\Sigma^p_{\mathcal{S}^p_{\mathcal{G}}\mathcal{S}^p_{\mathcal{G}}})^{-1}\left[-n^{-1}(Z^p_{\mathcal{S}^p_{\mathcal{G}}})^\top\epsilon + \lambda_1\left(\hat{B}_{\mathcal{S}^p_{\mathcal{G}}}\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}} - \hat{b}_p\mathbf{1}_{s-1}\right) + \lambda_2\left\{\operatorname{sgn}(\hat{\beta}_{\mathcal{S}^p_{\mathcal{G}}}) - \operatorname{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\right\}\right]\right.\\
&\qquad \left. + n^{-1}(Z^p_{\mathcal{S}^c_j})^\top\epsilon + \lambda_2\operatorname{sgn}(\beta_p)\mathbf{1}_{p_j} + \lambda_1 b_q\beta_p\mathbf{1}_{p_j}\right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\| n^{-1}(Z^p_{\mathcal{S}^c_j})^\top \epsilon - n^{-1}\Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top \epsilon \right\|_2 \\
&\quad + \lambda_1 \left\| \Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(\hat{B}_{\mathcal{S}^p_\mathcal{G}}\hat{\beta}_{\mathcal{S}^p_\mathcal{G}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}) + b_q\beta_p\mathbf{1}_{p_j} \right\|_2 \\
&\quad + \lambda_2 \left\| \Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1} \left\{ \mathrm{sgn}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}) - \mathrm{sgn}(\hat{\beta}_p)\mathbf{1}_s \right\} + \mathrm{sgn}(\beta_p)\mathbf{1}_{p_j} \right\|_2
\end{aligned}
$$

and so, for $V = \{V_j : j \in \mathcal{G}^c\}$, we have

$$
\begin{aligned}
\|V\|_\infty &\leq \max_{j\in\mathcal{G}^c} \left\| n^{-1}(Z^p_{\mathcal{S}^c_j})^\top \epsilon - n^{-1}\Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top \epsilon \right\|_2 \\
&\quad + \lambda_1 \max_{j\in\mathcal{G}^c} \left\| \Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(\hat{B}_{\mathcal{S}^p_\mathcal{G}}\hat{\beta}_{\mathcal{S}^p_\mathcal{G}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}) + b_q\beta_p\mathbf{1}_{p_j} \right\|_2 \\
&\quad + \lambda_2 \max_{j\in\mathcal{G}^c} \left\| \Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1} \left\{ \mathrm{sgn}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}) - \mathrm{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1} \right\} + \mathrm{sgn}(\beta_p)\mathbf{1}_{p_j} \right\|_2 \\
&\leq \sqrt{\kappa} \left\| n^{-1}(Z^p_{\mathcal{S}^c_j})^\top \epsilon - n^{-1}\Sigma^p_{\mathcal{S}^c_j\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(Z^p_{\mathcal{S}^p_\mathcal{G}})^\top \epsilon \right\|_\infty \\
&\quad + \lambda_1\sqrt{\kappa} \left\| \Sigma^p_{\mathcal{S}^c_{\mathcal{G}^c}\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}(B_{\mathcal{S}^p_\mathcal{G}}\hat{\beta}_{\mathcal{S}_{\setminus p}} - \hat{b}_q\hat{\beta}_p\mathbf{1}_{s-1}) + b_q\beta_p\mathbf{1}_{p_j} \right\|_\infty \\
&\quad + \lambda_2\sqrt{\kappa} \left\| \Sigma^p_{\mathcal{S}^c_{\mathcal{G}^c}\mathcal{S}^p_\mathcal{G}}(\Sigma^p_{\mathcal{S}^p_\mathcal{G}\mathcal{S}^p_\mathcal{G}})^{-1}\{\mathrm{sgn}(\hat{\beta}_{\mathcal{S}^p_\mathcal{G}}) - \mathrm{sgn}(\hat{\beta}_p)\mathbf{1}_{s-1}\} + \mathrm{sgn}(\beta_p)\mathbf{1}_{p_j} \right\|_\infty \\
&\leq \sqrt{\kappa} \left\{ \frac{\min(\lambda_1,\lambda_2)\eta}{\sqrt{\kappa}} \right\} + \frac{\lambda_1(1-\eta)}{2} \left\{ \frac{\min(\lambda_1,\lambda_2)}{\max(\lambda_1,\lambda_2)} \right\} + \frac{\lambda_2(1-\eta)}{2} \left\{ \frac{\min(\lambda_1,\lambda_2)}{\max(\lambda_1,\lambda_2)} \right\} \\
&\leq \lambda_1\eta + \lambda_1(1-\eta) = \lambda_1
\end{aligned}
$$

by (P2), Conditions (C1a) and (C1b), and the triangle inequality. This verifies (A3) and completes the proof.

C.4 Additional Simulation Results

In this section, we present mean squared prediction error across a variety of simulation settings for the sparse group lasso with $\theta = 0.05$ (SGL05) and with $\theta = 0.95$ (SGL95), the corresponding compositional sparse group lasso (CSGL05 and CSGL95), ridge regression, and kernel-penalized regression, as described in the main text. The group lasso and compositional group lasso perform very similarly to SGL05 and CSGL05, and similarly, the lasso and compositional lasso have prediction error very similar to SGL95 and CSGL95. These methods are therefore excluded from the plots for clarity.

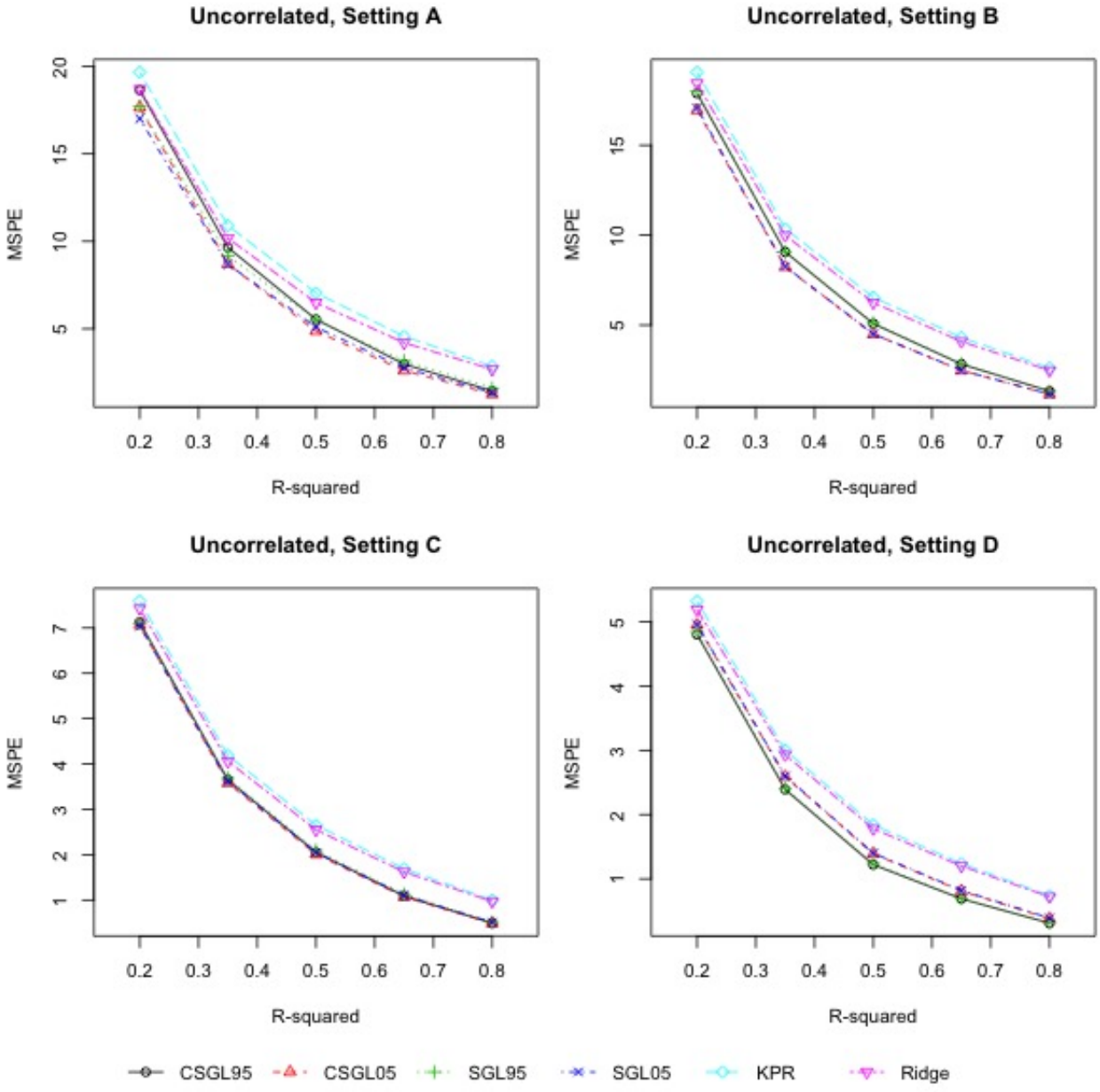


Figure C.1: Prediction error with uncorrelated taxa.

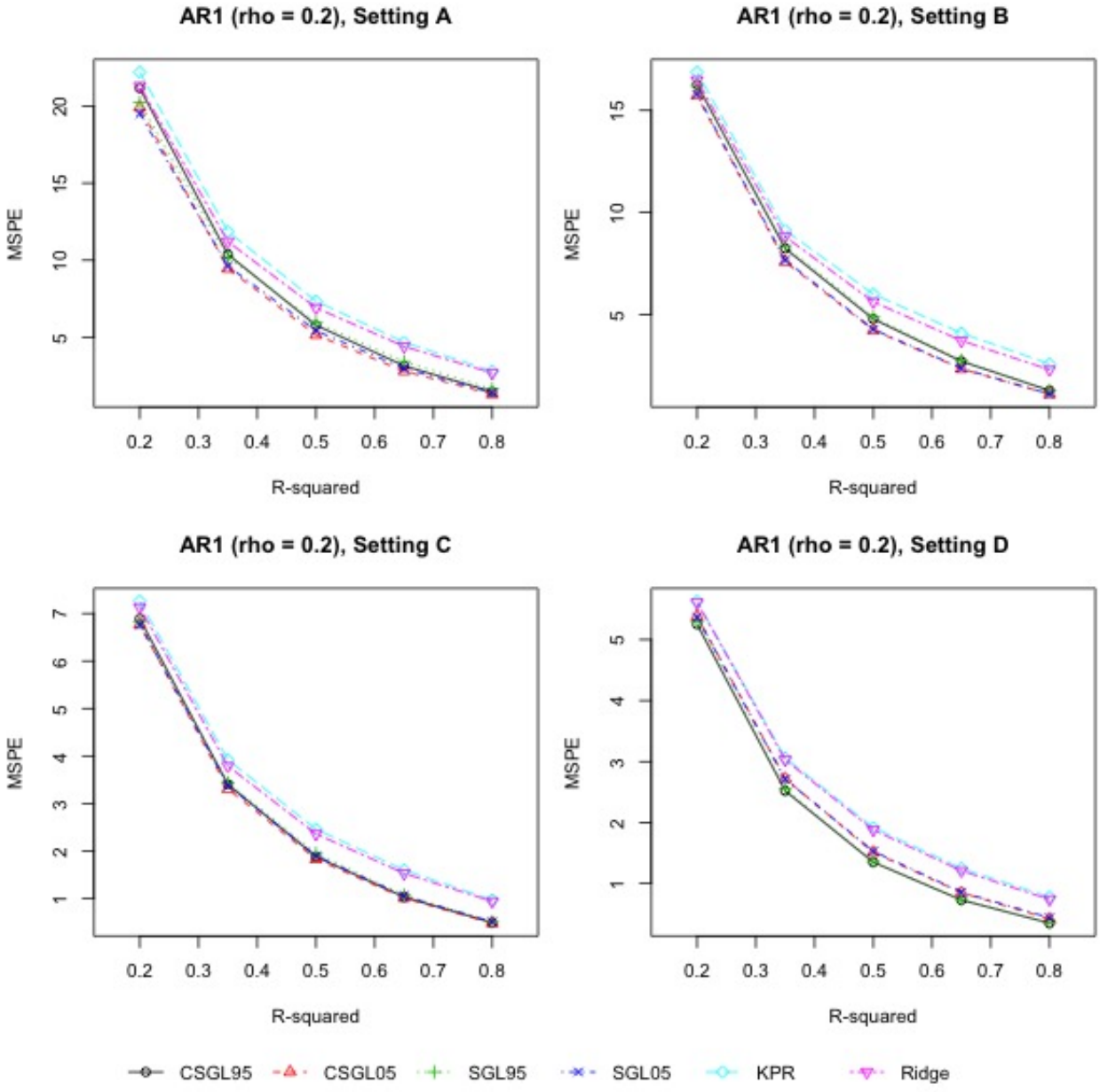


Figure C.2: Prediction error with AR(1) correlation structure among taxa, with $\rho = 0.2$.

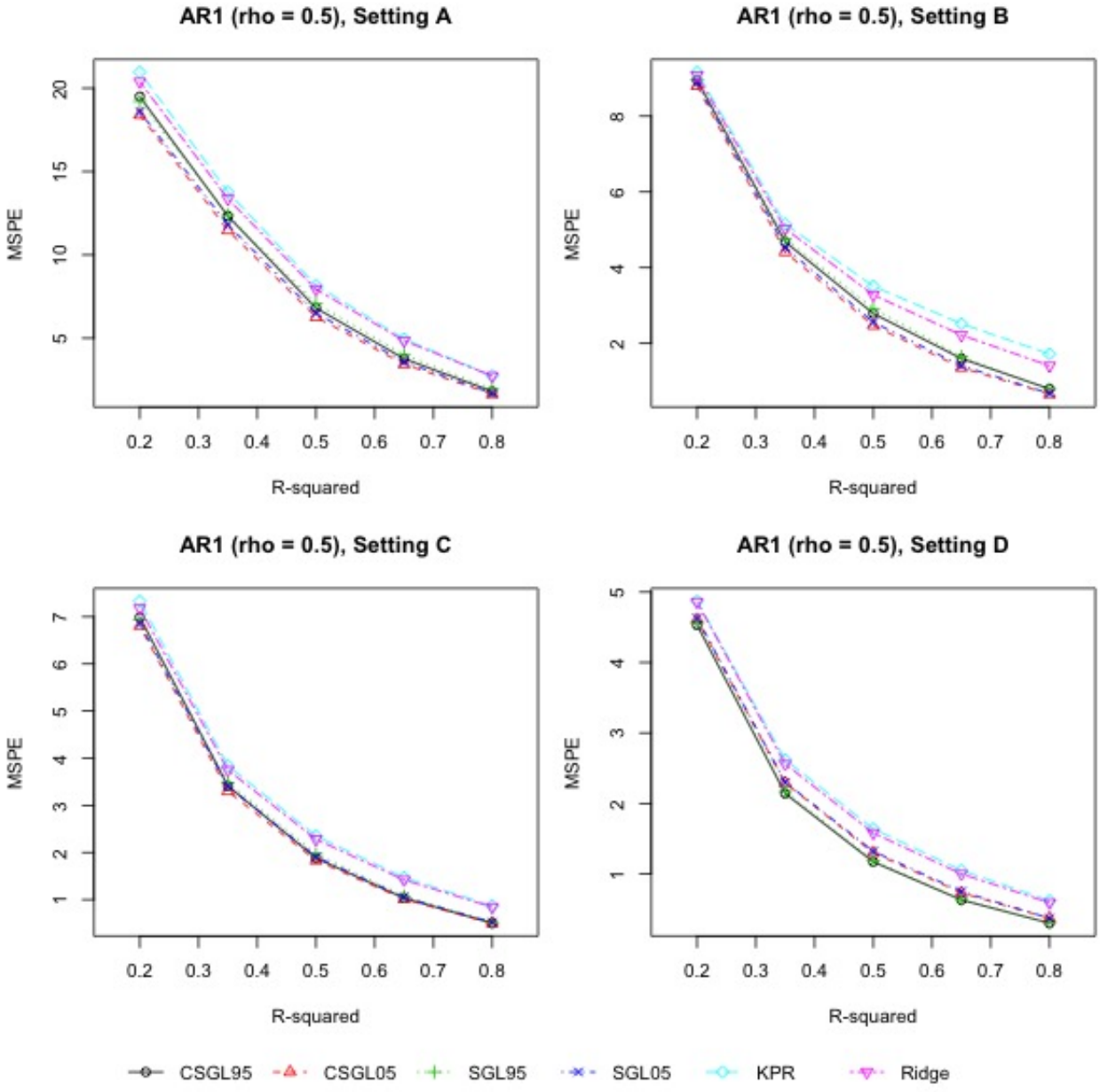


Figure C.3: Prediction error with AR(1) correlation structure among taxa, with $\rho = 0.5$.

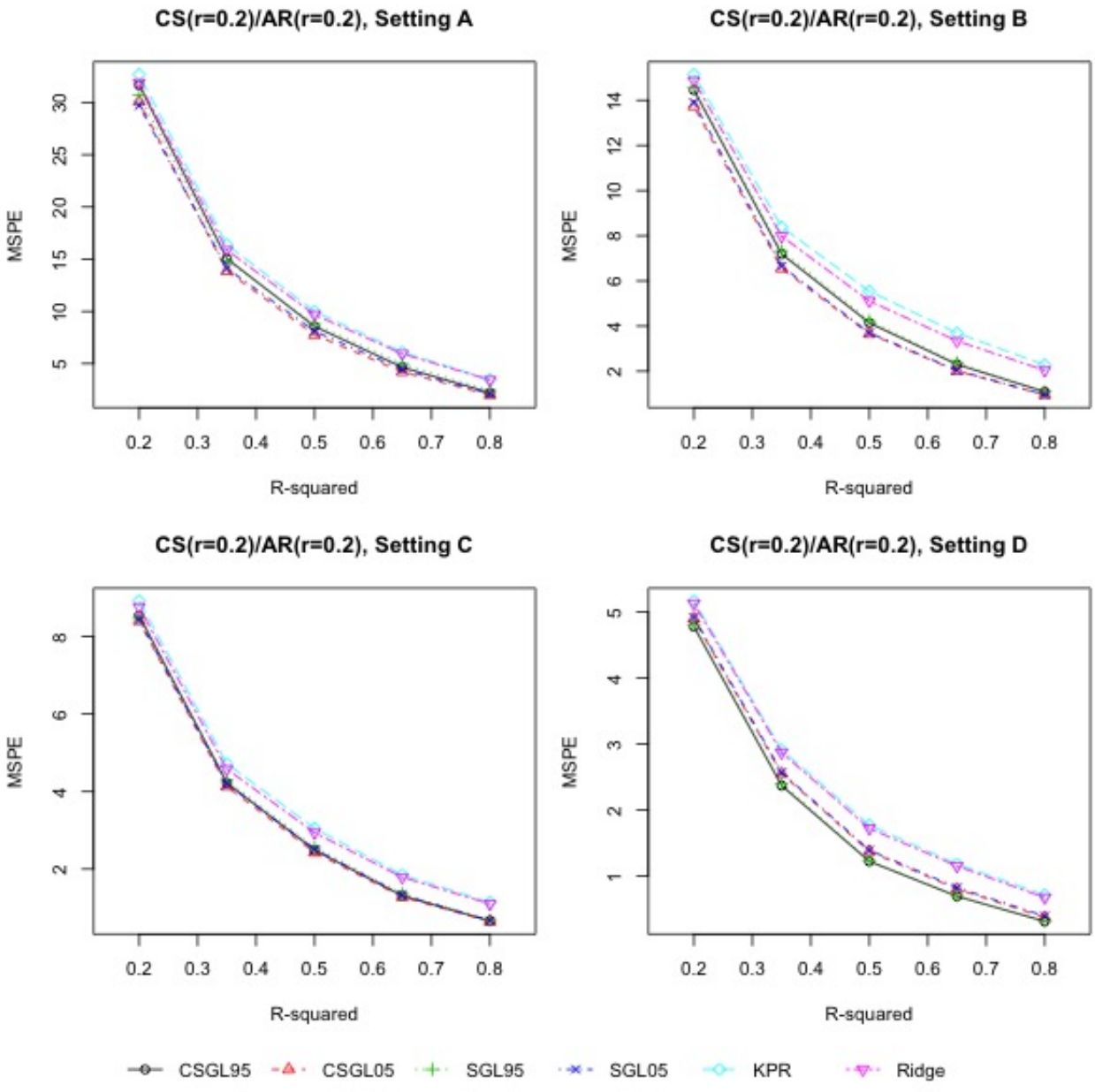


Figure C.4: Prediction error with compound symmetric correlation structure within groups ($\rho = 0.2$) and AR(1) correlation structure otherwise ($\rho = 0.2$).

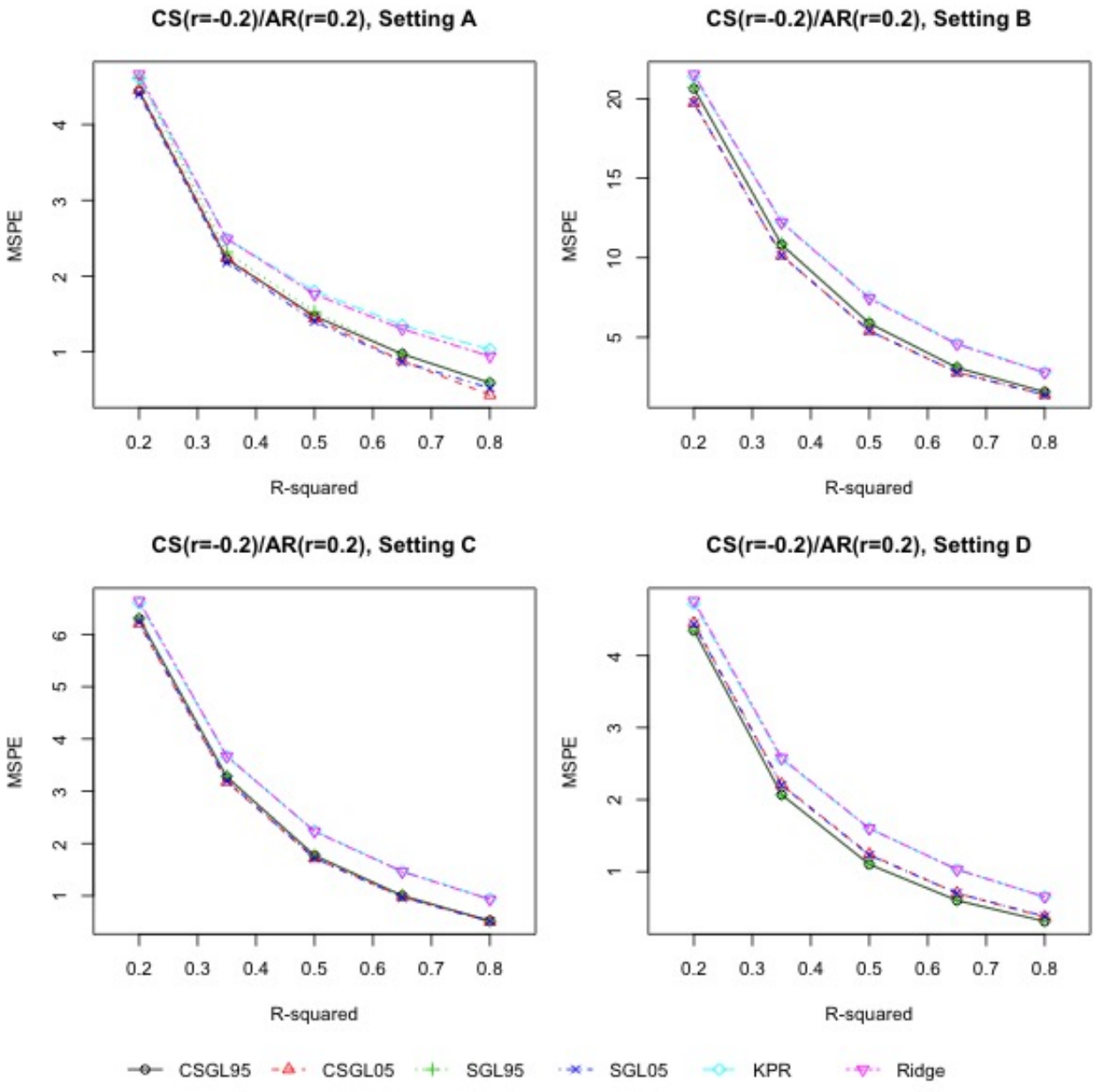


Figure C.5: Prediction error with compound symmetric correlation structure within groups ($\rho = -0.2$) and AR(1) correlation structure otherwise ($\rho = 0.2$).

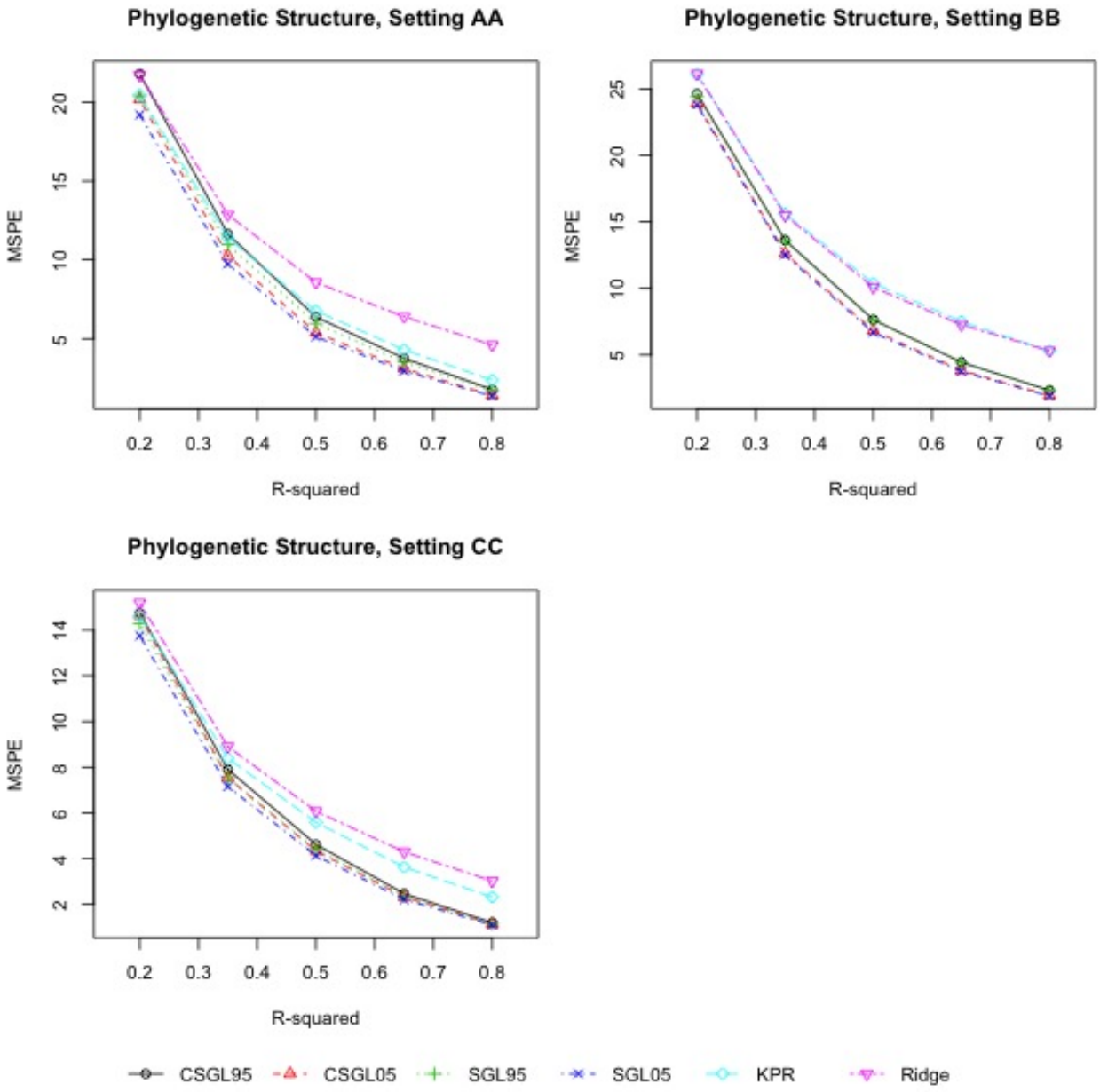


Figure C.6: Prediction error with features simulated using a Dirichlet-multinomial distribution with parameters estimated from a real microbiome dataset, with groups defined by phylogenetic relationships among taxa.