Title

A Semiparametric Model for Between-Subject Attributes: Applications to Beta-Diversity of Microbiome Data

Journal
Biometrics, 78(3)

ISSN
0006-341X

Authors
Liu, J
Zhang, Xinlian
Chen, T
et al.

Publication Date
2022-09-01

DOI
10.1111/biom.13487

Peer reviewed

Published in final edited form as:
Biometrics. 2022 September ; 78(3): 950–962. doi:10.1111/biom.13487.

A Semiparametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data

J. Liu[1,9], **X. Zhang**[1,*], **T. Chen**[2], **T. Wu**[1,9], **T. Lin**[1], **L. Jiang**[1,12], **S. Lang**[3], **L. Liu**[1], **L. Natarajan**[1], **J.X. Tu**[4], **T. Kosciolek**[5,6], **J. Morton**[7], **T.T Nguyen**[8,9], **B. Schnabl**[3], **R. Knight**[5,10,11,12], **C. Feng**[13], **Y. Zhong**[14], **X.M. Tu**[1,9]

[1]Department of Family Medicine and Public Health, UC San Diego, San Diego, California, U.S.A.

[2]Department of Mathematics, University of Toledo, Toledo, Ohio, U.S.A.

[3]Department of Medicine, UC San Diego, San Diego, California, U.S.A.

[4]Physical Medicine and Rehabilitation, University of Virginia Health System, Charlottesville, Virginia, U.S.A.

[5]Department of Pediatrics, UC San Diego, San Diego, California, U.S.A.

[6]Ma lopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

[7]Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, New York, U.S.A.

[8]Department of Psychiatry, UC San Diego, San Diego, California, U.S.A.

[9]Stein Institute for Research on Aging, UC San Diego, San Diego, California, U.S.A.

[10]Department of Computer Science and Engineering, UC San Diego, San Diego, California, U.S.A.

[11]Department of Bioengineering, UC San Diego, San Diego, California, U.S.A.

[12]Center for Microbiome Innovation, UC San Diego, San Diego, California, U.S.A.

[13]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.

[14]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

Summary:

The human microbiome plays an important role in our health and identifying factors associated with microbiome composition provides insights into inherent disease mechanisms. By amplifying and sequencing the marker genes in high-throughput sequencing, with highly similar sequences binned together, we obtain Operational Taxonomic Units (OTU) profiles for each subject. Due

[*] xizhang@health.ucsd.edu .

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2–4 as well as a zip file of code are available with this paper at the Biometrics website on Wiley Online Library. The R and Rcpp code can also be found at https://github.com/Jinyuan03140314/MicrobiomeFRM.

to the high-dimensionality and non-normality features of the OTUs, the measure of diversity is introduced as a summarization at the microbial community level, including the distance-based Beta-diversity between individuals. Analyses of such between-subject attributes are not amenable to the predominant within-subject based statistical paradigm, such as t-tests and linear regression. In this paper, we propose a new approach to model Beta-diversity as a response within a regression setting by utilizing the functional response models (FRM), a class of semiparametric models for between- as well as within-subject attributes. The new approach not only addresses limitations of current methods for Beta-diversity with cross-sectional data, but also provides a premise for extending the approach to longitudinal and other clustered data in the future. The proposed approach is illustrated with both real and simulated data.

Keywords

Copula; Functional Response Model; High-throughput Sequencing; Permutational Multivariate Analysis of Variance Using Distance Matrices (PERMANOVA); Semiparametric Regression; U-statistics based Generalized Estimating Equation (UGEE)

1. Introduction

This methodological development is motivated by the problem to test associations between the microbiome diversity and clinical variables. The human microbiome refers to all microorganisms on or in the human body, their genes, and surrounding environmental conditions (National Academies of Sciences and Medicine, 2018). In recent years, a preponderance of microbiome studies have implicated the role of the human microbiome in the pathogenesis of complex diseases, including diabetes, alcoholic liver disease, and even cancers (Lang et al., 2020b; Holmes et al., 2011). Therefore, identifying potential biological or clinical variables associated with the microbiome and defining their relationships not only enlighten the inherent disease mechanisms but also enhance modulating microbiome compositions for therapeutic purposes.

Fueled by the technological advancement of next-generation sequencing, the human microbiome can be interrogated using high-throughput sequencing. For example, one strategy amplifies and sequences the bacterial 16S ribosomal RNA gene (16S rRNA) for species identification. These sequences are further clustered into nearly identical Operational Taxonomic Units (OTUs) and compared with reference databases to produce OTU counts profiles based on taxonomic assignments. The OTU counts are often sparse and high-dimensional. Direct analysis of such data with limited samples raises several statistical challenges, including modeling the skewed and over-dispersed count data with a preponderance of zeros. Since the sequencing depth varies, OTU counts are usually normalized into proportions within each subject to form the OTU relative abundance. They can be further summarized at the microbial community level using diversity metrics, including the "within-subject" Alpha-diversity and "between-subject" Beta-diversity. Unlike Alpha-diversity that consists of individual outcomes, or within-subject attributes, Beta-diversity considers the number of shared taxa between subjects, thus representing their differences in OTU abundance profiles. Each Beta-diversity outcome is a pairwise distance between two subjects, or between-subject attribute. The two major categories of statistical analyses for the microbiome, i.e., the "individual" level effect of a single OTU and the "community" level effect of microbiome composition with summary statistics of diversity, complement each other.

Notably, a variety of disorders are shown to be associated with the loss of gut microbial diversity (Durack and Lynch, 2019). One common approach to evaluate such associations using Beta-diversity is the Permutational Multivariate Analysis of Variance Using Distance Matrices (PERMANOVA) (McArdle and Anderson, 2001). This approach partitions the Beta-diversity into within- and between-group variations and implements a permutation test based on pseudo-F statistics for inference. A major limitation is the difficulty to discern the sources of variation when the null hypothesis is rejected. Also, it is unsuitable for between-subject covariates in some applications, such as a dissimilarity measure describing the difference between subjects' metabolites abundance profile. Additionally, it requires a large number of permutations to ensure stable results (Dubitzky et al., 2013). All these limitations severely circumscribe its applications in practice.

We propose a new approach to address the aforementioned limitations of PERMANOVA by utilizing the functional response models (FRM) (Kowalski and Tu, 2008), which are uniquely positioned to address between-subject attributes defining the Beta-diversity in the current context. In Section 2, we provide a brief overview of the Beta-diversity and PERMANOVA. In Section 3, we develop the proposed approach for Beta-diversity within a regression setting. In Section 4, we first develop a new approach to simulate life-like OTU counts and Beta-diversity, and then evaluate performance of the proposed and existing approaches. We conclude this section with an application to a study on alcoholic liver disease. In Section 5, we give our concluding remarks.

2. Beta-diversity and PERMANOVA

2.1 Beta-diversity Measures

Beta-diversity captures within- and between-group differences by comparing individuals' distributions of taxonomic units. For example, the Bray-Curtis distance (Sørensen, 1948) is a quantitative measure based on OTU relative abundance. For a pair of subjects i and j, the Bray-Curtis distance is defined by $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$, where C_{ij} indicates the sum of the OTU relative abundance that the pair has in common and S_i (S_j) denotes the total number of OTU relative abundance for the ith (jth) subject. This measure ranges from 0 to 1, with 0 (1) indicating exactly the same (completely different) taxonomic abundances. As Beta-diversity incorporates taxa information into distances, its size is determined by the number of subjects rather than that of taxonomic units for the high-dimensional OTUs.

Unlike the Euclidean distance, most Beta-diversity measures calculate weighted relative differences, where each species' contribution is weighted by the sum of the species' abundance in the two subjects being compared (Roberts, 2017). Some forms such as the Unifrac can additionally account for the phylogenetic distances (Lozupone and Knight, 2005). Hence, non-Euclidean Beta-diversity measures are widely adopted as the basis of statistical analyses to detect a wider range of biologically relevant changes in the microbiome (Legendre and Gallagher, 2001).

2.2 PERMANOVA

Consider a sample of n subjects with microbiome profiles (counts) defined by m OTUs. Let $\mathbf{y}_i$ denote an $m \times 1$ column vector of OTU relative abundance (after normalization) and $\mathbf{x}_i$ a vector of explanatory variables such as the status of a disease for the ith subject. Let $d_{\mathbf{i}} = d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ denote a Beta-diversity outcome for a pair of subjects $\mathbf{i} = (i_1, i_2) \in C_2^n$, where C_q^n denotes the set of q-combinations $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$.We are interested in testing the association between the Beta-diversity $d_{\mathbf{i}}$ and some clinical variables such as the status of a disease or, more generally, a continuous explanatory variable such as bilirubin, an indication of liver disease progression.

If $\mathbf{x}_i$ is a categorical variable for groups, PERMANOVA can be used to compare Beta-diversity across different groups, which adopts a pseudo-F statistic for inference (McArdle and Anderson, 2001). We provide details and formulas in the Supporting Information.

PERMANOVA has several limitations. First, it does not provide coefficient estimators for explanatory variables, which hinders generating interpretable results on both the direction and size of the effects, or discerning sources of differences. Second, it describes relationships of Beta-diversity (a between-subject attribute) with within-subject attributes only, not between-subject attributes such as metabolites abundance profile. Also, it requires a large number of permutations for stable results and thus carries more overheads in terms of the computational burden. Additionally, it is quite difficult to extend PERMANOVA to longitudinal studies (with missing data) that are potentially valuable given the dynamic and highly personalized nature of the microbiome.

3. Functional Response Models for Beta-diversity

The aforementioned limitations of PERMANOVA result from a lack of ability to model between-subject attributes under the predominant statistical paradigm. With a few exceptions such as the Mann-Whitney-Wilcoxon rank-sum test (Wu et al., 2014; Lin et al., 2021), all popular statistical models focus on relationships between variables from the same subject, or within-subject attributes. As Beta-diversity measures the difference between a pair of subjects' OTUs, conventional statistical models are not amenable to modeling such between-subject attributes. In this section, we develop a regression framework to model Beta-diversity by utilizing a class of functional response models (FRM).

3.1 Functional Response Models for Between-subject Attributes

Consider a class of semiparametric functional response models (FRM):

$$E\{\mathbf{f}(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q}) \mid \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}\} = \mathbf{h}\left(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}; \theta\right), \quad \left(i_1, \ldots, i_q\right) \in C_q^n, \quad 1 \leqslant q, \quad 1 \leqslant i \leqslant n, \tag{1}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top \in \mathbb{R}^m$ denotes the response vector from the ith subject, $\mathbf{f}(\cdot)$ is some vector-valued function, $\mathbf{h}(\cdot)$ is some vector-valued smooth function (e.g., with continuous derivatives up to the second order), $\boldsymbol{\theta}$ is a vector of parameters, q is some positive integer. The FRM in (1) extends the semiparametric generalized linear models (GLM) from within- to between-subject attributes (Kowalski and Tu, 2008). For example, when $q = 1$ and $f(y_i) = y_i$, (1) immediately reduces to the restricted moment GLM. When $q = 2$ and set

$$f_{\mathbf{i}} = d\left(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\right), \quad h_{\mathbf{i}}(\theta) = E\left\{d\left(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\right)\right\} = \theta, \quad (i_1, i_2) \in C_2^n, \tag{2}$$

the FRM in (1) models the Beta-diversity distance $d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ and provides inference about the mean distance θ.

3.2 Functional Response Models for Beta-diversity with Covariates

3.2.1 Group Comparison.—We start by comparing Beta-diversity across multiple groups. Consider K groups with n_k denoting the sample size of the kth group $(1 \leqslant k \leqslant K)$, $n = \sum_{k=1}^{K} n_k$ denoting the total sample size of all K groups combined. Let x_i denote a categorical variable indicating group membership for subject $i(1 \leqslant x_i \leqslant K,\ 1 \leqslant i \leqslant n)$.

For each pair, we observe their OTU relative abundance outcomes $\mathbf{y_i} = \left\{\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\right\}$ $\left(\mathbf{i} = (i_1, i_2) \in C_2^n\right)$ along with the pairwise group indicators $\mathbf{x_i} = \left\{x_{i_1}, x_{i_2}\right\}\left(1 \leqslant x_{i_1}, x_{i_2} \leqslant K\right)$. Denote all combinations of $\mathbf{x_i}$ with a vector $\boldsymbol{\delta}(\mathbf{x_i}) \in \mathbb{R}^{K + C_2^K}$ through a one-hot encoding function $\delta: \{1, \ldots, K\} \times \{1, \ldots, K\} \mapsto \{0, 1\}^{K + C_2^K}$ such that for its $\mathbf{k}^{th}$ ($\mathbf{k} = \{k_1, k_2\}$) entry:

$$\begin{aligned} \delta_{\mathbf{k}}(\mathbf{x_i}) &= \begin{cases} 1 & \text{if } \mathbf{x_i} = \left\{x_{i_1}, x_{i_2}\right\} = \{k_1, k_2\} = \mathbf{k} \\ 0 & \text{otherwise} \end{cases}, \quad \mathbf{i} = (i_1, i_2) \in C_2^n, \\ \boldsymbol{\delta}(\mathbf{x_i}) &= \left(\delta_{11}(x_{\mathbf{i}}), \ldots, \delta_{(K-1)K}(x_{\mathbf{i}}), \delta_{KK}(x_{\mathbf{i}})\right)^\top, \quad 1 \leqslant k_1 \leqslant k_2 \leqslant K\,. \end{aligned} \tag{3}$$

Let $f(\mathbf{y_i}) = d\left(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\right)$ and define an FRM:

$$E\{f(\mathbf{y_i}) \mid \boldsymbol{\delta}(\mathbf{x_i})\} = \exp\left\{\sum_{1 \leqslant k_1 \leqslant k_2 \leqslant K} \tau_{k_1 k_2} \delta_{k_1 k_2}(\mathbf{x_i})\right\} = \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\delta}(\mathbf{x_i})\right\}, \tag{4}$$

where exp $(\cdot)$ ensures that the right side of the equation is positive as $f(\mathbf{y_i}) \geqslant 0$. The FRM above is determined by the parameter vector $\boldsymbol{\theta} = \left(\tau_{11}, \ldots, \tau_{(K-1)K}, \tau_{KK}\right)^\top$.

Unlike conventional analysis for within-subject attributes, models for between-subject attributes involve more complex parameters and interpretations. For the FRM in (4), $\exp(\tau_{kk})$ is the mean of $f(\mathbf{y_i})$ when both subjects of the $\mathbf{i}^{th}$ pair are from group k, and $\exp(\tau_{k_1 k_2})$ is the mean of $f(\mathbf{y_i})$ when one (the other) is from group k_1 (k_2). Thus, in addition to group means as in conventional within-subject analysis, we now have (1) within-group means

$\exp(\tau_{kk})$ and (2) between-group means $\exp(\tau_{k_1k_2})$ For two groups k_1 and k_2 with the same or similar OTU distributions, their within- and between-group means are usually similar. However, if they have different OTU distributions, they may still have similar within-group means (this can occur, for example, if OTUs' have similar variability within each group), but the between-group means $\exp(\tau_{k_1k_2})$ can be different from within-group means $\exp(\tau_{k_1k_1})$ or $\exp(\tau_{k_2k_2})$.

Thus, under the FRM in (4), we are interested in three types of null hypotheses to describe group differences in Beta-diversity:

$$
\begin{aligned}
&\text{(1) Within-group}: \begin{cases} H_{01}: \tau_{kk} = \tau_{k'k'} & \text{for any}(k, k'), 1 \leqslant k < k' \leqslant K \\ H_{a1}: \tau_{kk} \neq \tau_{k'k'} & \text{for some}(k, k') \end{cases}, \\
&\text{(2) Between-group}: \begin{cases} H_{02}: \tau_{kl} = \tau_{k'l'} & \text{for any}(k, l, k', l'), 1 \leqslant k, k' < l, l' \leqslant K \\ H_{a2}: \tau_{kl} \neq \tau_{k'l'} & \text{for some}(k, l, k', l') \end{cases} \\
&\text{(3) Within- vs. Between-group}: \begin{cases} H_{03}: \tau_{kk} = \tau_{k'l'} & \text{for any}(k, k', l'), 1 \leqslant k \leqslant K,\ 1 \leqslant k' < l' \leqslant K \\ H_{a3}: \tau_{kk} \neq \tau_{k'l'} & \text{for some}(k, k', l') \end{cases}.
\end{aligned}
\tag{5}
$$

Hypotheses (2) and (3) are unique to between-subject attributes, each revealing different aspects. For example, if the patterns of OTU distribution are "flipped" across two groups, the difference of Beta-diversity could be detected by the "within- vs. between-" instead of the "within-" type of hypothesis.

For PERMANOVA, if we obtain an insignificant pseudo-F statistic, we conclude with not enough evidence to reject the null. But, if this test is significant, it is unclear if the difference occurs in within-group or between-group means or both. By partitioning sources of variation and building formal hypotheses to depict the underlying differences of microbiome diversity across groups, a formal regression model for between-subject attributes in (4) allows for discerning sources of differences, potentially leading to more in-depth scientific findings.

All three types of hypotheses in (5) are readily tested using linear contrasts: $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{0}$ vs. $H_a : \mathbf{C}\boldsymbol{\theta} \neq \mathbf{0}$, where $\mathbf{C}$ is a matrix of known constants. For example, when comparing Beta-diversity for three groups, we may use the following $\mathbf{C}$ matrices to test the hypotheses in (5):

$$
\begin{aligned}
&K = 3,\ \boldsymbol{\theta} = (\tau_{11}, \tau_{22}, \tau_{33}, \tau_{12}, \tau_{13}, \tau_{23})^{\top}, \quad (a): \mathbf{C}_1 = \left(\mathbf{1}_2,\ (-1) \cdot \mathbf{I}_2,\ 0_{2 \times 3}\right); \\
&(b): \mathbf{C}_2 = \left(0_{2 \times 3},\ \mathbf{1}_2,\ (-1) \cdot \mathbf{I}_2\right); \quad (c): \mathbf{C}_3 = \left(\mathbf{1}_5,\ (-1) \cdot \mathbf{I}_5\right),
\end{aligned}
\tag{6}
$$

where $\mathbf{1}_n$ denotes a $n \times 1$ column vector of 1's, and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

3.2.2 Covariates for Confounders.

—As most human population studies of microbiome are observational due to cost, logistic, and difficulties in experimental control, it is crucial to control for potential confounders that may impact group differences, such as demographics (ethnicity, genetic background), biometrics (medications, diet), molecular measures (microbial metabolites, gene expression), and environmental exposures (National Academies of Sciences and Medicine, 2018). A more substantial improvement over PERMANOVA is FRM's ease to control for a broader range of confounders, including between-subject attributes such as metabolites abundance profiles. This is achieved by leveraging the regression feature of FRM to include either within- or between-subject covariates.

As a motivating example for including a within-subject covariate, consider a linear regression relating a continuous variable z_i to a continuous response y_i: $y_i = \eta_0 + \eta_1 z_i + \epsilon_i$, $\epsilon_i \sim (0, \sigma^2)$, $1 \leqslant i \leqslant n$, where $(0, \sigma^2)$ denotes some continuous distribution with mean zero and variance σ^2. Now consider the squared difference, $f(y_{\mathbf{i}}) = \left(y_{i_1} - y_{i_2}\right)^2$. It follows that

$$E\left\{f(y_{\mathbf{i}}) \middle| z_{i_1}, z_{i_2}\right\} = E\left(\epsilon_{i_1} - \epsilon_{i_2}\right)^2 + \eta_1^2\left(z_{i_1} - z_{i_2}\right)^2 = 2\sigma^2 + \eta_1^2\left(z_{i_1} - z_{i_2}\right)^2. \tag{7}$$

Although Beta-diversity is more complex, we use the same rationale to control for covariates by adding $\left(z_{i_1} - z_{i_2}\right)^2$, or a more general non-negative transformation $g(\mathbf{z_i})$ of $\mathbf{z_i} = \left\{z_{i_1}, z_{i_2}\right\}$ to the FRM in (4):

$$E\{f(y_{\mathbf{i}}) \mid \boldsymbol{\delta}(\mathbf{x_i}), \mathbf{z_i}\} = \exp\left\{\sum_{1 \leqslant k_1 \leqslant k_2 \leqslant K} \tau_{k_1 k_2} \delta_{k_1 k_2}(\mathbf{x_i}) + \xi_1 g(\mathbf{z_i})\right\}, \ \mathbf{i} = (i_1, i_2) \in C_2^n. \tag{8}$$

For a categorical covariate, we can define a series of indicators akin to (3), i.e. for the $\mathbf{i}^{th}$ pair, we observe the pairwise indicators $\mathbf{x}_{l\mathbf{i}} = \left\{x_{li_1}, x_{li_2}\right\} \left(1 \leqslant x_{li_1}, x_{li_2} \leqslant K_l\right)$ for the l^{th} $(1 \leqslant l \leqslant p)$ categorical covariate with K_l levels. We one-hot encode those p categorical covariates into $\boldsymbol{\delta}(\mathbf{x_i}) \in \mathbb{R}^{1 + \sum_{l=1}^{p}\left(K_l + C_2^{K_l} - 1\right)}$, with the encoding function defined similarly as in (3), but designating a referent to obtain a similar form as in conventional regression.

Specifically, for the l^{th} categorical covariate, we define $\delta_l: \{1, \ldots, K_l\} \times \{1, \ldots, K_l\} \mapsto \{0, 1\}^{K_l + C_2^{K_l} - 1}$ (excluding the case where $k_{l1} = k_{l2} = 1$)such that for the $\mathbf{k}_l^{th}(\mathbf{k}_l = \{k_{l1}, k_{l2}\})$ entry of $\boldsymbol{\delta}_l(\mathbf{x}_{l\mathbf{i}})$:

$$\begin{aligned} \delta_{l\mathbf{k}}(\mathbf{x}_{l\mathbf{i}}) &= \begin{cases} 1 \ \text{ if } \mathbf{x}_{l\mathbf{i}} = \left\{x_{li_1}, x_{li_2}\right\} = \{k_{l1}, k_{l2}\} \quad = \mathbf{k}_l \\ 0 \ \text{ otherwise} \end{cases}, \\ \boldsymbol{\delta}_l(\mathbf{x}_{l\mathbf{i}}) &= \left(\delta_{l12}(\mathbf{x}_{l\mathbf{i}}), \ldots, \delta_{l(K-1)K}(\mathbf{x}_{l\mathbf{i}}), \delta_{lKK}(\mathbf{x}_{l\mathbf{i}})\right)^\top, \quad 1 \leqslant l \leqslant p, \\ \boldsymbol{\delta}(\mathbf{x_i}) &= \left(1, \boldsymbol{\delta}_1(\mathbf{x}_{1\mathbf{i}})^\top, \ldots, \boldsymbol{\delta}_l(\mathbf{x}_{l\mathbf{i}})^\top, \ldots, \boldsymbol{\delta}_p(\mathbf{x}_{p\mathbf{i}})^\top\right)^\top, \\ \mathbf{i} &= (i_1, i_2) \in C_2^n, \quad 1 \leqslant k_{l1} \leqslant k_{l2} \leqslant K_l, \quad 1 = k_{l1} \neq k_{l2}. \end{aligned} \tag{9}$$

Thus, with p categorical covariates (including the one for diagnostic groups), $x_{li}(1 \leqslant l \leqslant p)$, and q continuous covariates, $z_{mi}(1 \leqslant m \leqslant q)$ for subject i, we can, after designating the first group as the referent by including an intercept β_0, express the FRM as:

$$E\{f(\mathbf{y_i})|\mathbf{x_i}, \mathbf{z_i}\} = \exp\left\{\beta_0 + \sum_{l=1}^{p}\left(\sum_{1 \leqslant k_{l1} \leqslant k_{l2} \leqslant K_l}^{1 = k_{l1} \neq k_{l2}} \beta_{lk_1k_2}\delta_{lk_1k_2}(\mathbf{x_{li}})\right) + \sum_{m=1}^{q} \xi_m g_m(\mathbf{z_{mi}})\right\}, = \exp\left\{\boldsymbol{\beta}^\top \boldsymbol{\delta}(\mathbf{x_i}) + \boldsymbol{\xi}^\top \mathbf{g}(\mathbf{z_i})\right\}, \tag{10}$$

where $\mathbf{x_{li}} = \left\{x_{li_1}, x_{li_2}\right\}$, $\mathbf{z}_{mi} = \left\{z_{mi_1}, z_{mi_2}\right\}$, $\mathbf{g}(\mathbf{z_i}) = \left(g_1(\mathbf{z_{1i}}), \ldots, g_q(\mathbf{z_{qi}})\right)^\top$ and K_l denotes the levels of category of the l^{th} categorical variable $x_{li}(1 \leqslant l \leqslant p)$. The FRM above is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^{1 + \Sigma_{l=1}^{p}\left(K_l + C_2^{K_l} - 1\right) + q}$:

$$\boldsymbol{\beta}_l = \left(\beta_{l12}\ldots, \beta_{l(K_l - 1)K_l}, \beta_{lK_lK_l}\right)^\top, \quad \boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top\right)^\top, \quad \boldsymbol{\xi} = \left(\xi_1, \ldots, \xi_q\right)^\top, \quad \boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \boldsymbol{\xi}^\top\right)^\top. \tag{11}$$

Akin to (4), the parameters for the covariates possess more complex interpretations. For a continuous covariate $\mathbf{z_{mi}}$, ξ_m represents change in the mean of log $\{f(\mathbf{y_i})\}$ per unit change in $g_m(\mathbf{z_{mi}})$. For a categorical one, say gender, we now have male-male, female-female, or male-female pairs. If we set male-male as the referent, coefficients for female-female and male-female pairs represent differences in the log of mean Beta-diversity when comparing the respective gender pair to the referent.

We illustrate this model with a relatively simple log-linear form in (10), yet the applicability of FRM is far beyond the assumed simple relationship. Like any regression model such as the GLM, more complex relationships such as higher-order terms and interactions can be specified as deemed appropriate. The FRM in (10) looks like a conventional (log-linear) regression model, except that i indexes pairs of, rather than, individual, subjects. This critical difference precludes applications of standard inference methods for regression models as we discuss next.

3.2.3 Inference.—As the response function $f_\mathbf{i} = f(\mathbf{y_i})$ of the FRM-based regression for Beta-diversity in (10) involves pairs of subjects, inferences about $\boldsymbol{\theta}$ must address the interlocking dependence of $f_\mathbf{i}$'s. Since this type of dependence structure is not addressed by standard methods such as the Generalized Estimating Equations (GEE), we develop inferences using a class of U-statistics based Generalized Estimating Equations (UGEE).

U-statistics based Generalized Estimating Equations.: Let

$$S_{\mathbf{i}}=f_{\mathbf{i}}-h_{\mathbf{i}},\ \ \mathbf{D}_{\mathbf{i}}=\frac{\partial}{\partial\boldsymbol{\theta}}h_{\mathbf{i}},\quad V_{\mathbf{i}}=Var(f_{\mathbf{i}}\mid\mathbf{x}_{\mathbf{i}},\mathbf{z}_{\mathbf{i}}),\ \ \mathbf{i}=(i_1,i_2)\in C_2^n, \tag{12}$$

in practice, $V_{\mathbf{i}}$ is generally unknown and substituted by a working variance such as $V_{\mathbf{i}}(h_{\mathbf{i}}) = h_{\mathbf{i}}$, as the form of FRM is similar to log-linear models for within-subject attributes. Thus, define the UGEE:

$$\mathbf{U}_n(\theta)=\sum_{\mathbf{i}\in\mathbf{C}_2^n}\mathbf{U}_{n,\mathbf{i}}=\sum_{\mathbf{i}\in C_2^n}\mathbf{D}_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}}=0, \tag{13}$$

where the estimates $\hat{\theta}$ are obtained through the Newton-Raphson method (see the Supporting Information for details).

Although similar in appearance, the UGEE above is not a sum of independent variables as in GEE (Tang, He, and Tu, 2012). Standard asymptotic methods such as the central limit theorem cannot be applied directly, but the theory of U-statistics is useful for addressing such interlocking dependence. For ease of reference, we summarize the asymptotic properties in the theorem below and provide a sketch of proof in the Supporting Information.

Theorem 1. Let

$$\begin{aligned}&\mathbf{v}_{i_1}=E\left(\mathbf{U}_{n,\mathbf{i}}\middle|\mathbf{y}_{i_1},\mathbf{x}_{i_1},\mathbf{z}_{i_1}\right),\quad \mathbf{B}=E\left(\mathbf{D}_{\mathbf{i}}V_{\mathbf{i}}^{-1}\mathbf{D}_{\mathbf{i}}^{\top}\right),\\ &\boldsymbol{\Sigma}_U=4Var\left(\mathbf{v}_{i_1}\right),\quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}}=\mathbf{B}^{-1}\boldsymbol{\Sigma}_U\mathbf{B}^{-1},\quad \mathbf{i}=(i_1,i_2)\in C_2^n.\end{aligned} \tag{14}$$

Then under mild regularity conditions,

a. $\hat{\theta}$ is consistent and asymptotically normal:

$$\sqrt{n}\left(\hat{\theta}-\theta\right)\to_d N(\mathbf{0},\boldsymbol{\Sigma}_{\theta}) \tag{15}$$

where $\to_d$ denotes convergence in distribution.

b. A consistent estimate of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is obtained by substituting consistent estimates of $\boldsymbol{\theta}$ and moments of the respective quantities in $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

Theorem 1 above is readily applied to test any linear hypotheses concerning $\boldsymbol{\theta}$, such as the linear contrasts in (6). Under the null, the Wald statistic has an asymptotic χ^2 distribution:

$$W_n=n\left(\mathbf{C}\hat{\theta}\right)^{\top}\left(\mathbf{C}\widehat{\boldsymbol{\Sigma}}_{\theta}\mathbf{C}^{\top}\right)^{-1}\left(\mathbf{C}\hat{\theta}\right)\to_d\chi_s^2, \tag{16}$$

where s is the rank of $\mathbf{C}$ and χ_s^2 denotes a (central) χ^2 distribution with s degrees of freedom. For example, in testing the within-group difference H_{01} in (6), $W_n\to_d\chi_2^2$ under H_{01}.

The Score Test.: As Wald-type tests are typically anti-conservative, score statistics may be used as an alternative to reduce such bias, especially for small to moderate samples

(Kennedy, 2008). To develop a score statistic based on the UGEE in (13), let $\theta = \left(\theta_{(1)}^{\top}, \theta_{(2)}^{\top}\right)^{\top}$, where $\boldsymbol{\theta}_{(2)}$ is the parameter of interest, $\theta_{(1)} \in \mathbb{R}^{p}, \theta_{(2)} \in \mathbb{R}^{q}$. Consider testing the null $H_0 : \boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{(20)}$, with $\boldsymbol{\theta}_{(20)}$ a vector of known constants. We have the partition:

$$\mathbf{D_i} = \left(\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(1)}}, \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(2)}}\right)^{\top} = \left(\mathbf{D}_{\mathbf{i}(1)}, \mathbf{D}_{\mathbf{i}(2)}\right)^{\top}, \quad \mathbf{U}_n(\boldsymbol{\theta}) = \left(\mathbf{U}_{n(1)}(\boldsymbol{\theta}), \mathbf{U}_{n(2)}(\boldsymbol{\theta})\right)^{\top}, \tag{17}$$

let $\tilde{\theta}_{(1)}$ denote the estimate of $\boldsymbol{\theta}_{(1)}$ from solving the following reduced estimating equation given $\boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{(20)}$

$$\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right) = \binom{n}{2}^{-1} \sum_{i \in C_2^n} \mathbf{D}_{\mathbf{i}(1)} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = 0\,. \tag{18}$$

To define the score statistic, let

$$\begin{aligned} &\tilde{\theta} = \left(\tilde{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right)^{\top}, \quad \mathbf{B} = E\left(\mathbf{D_i} V_{\mathbf{i}}^{-1} \mathbf{D}_{\mathbf{i}}^{\top}\right) = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{12}^{\top} & \mathbf{B}_{22} \end{pmatrix}, \\ &\mathbf{G} = \left(-\mathbf{B}_{21}\mathbf{B}_{11}^{-1}, \ \mathbf{I}_q\right), \quad \boldsymbol{\Sigma}_{(2)} = \mathbf{G}\boldsymbol{\Sigma}_U \mathbf{G}^{\top}, \end{aligned} \tag{19}$$

where $\mathbf{I}_\mathrm{q}$ denotes the $q \times q$ identity matrix, $\mathbf{B}_{11} \in \mathbb{R}^{p \times p}$, $\mathbf{B}_{12} \in \mathbb{R}^{p \times q}$, and $\mathbf{B}_{22} \in \mathbb{R}^{q \times q}$ denote the respective submatrices from partitioning the matrix $\mathbf{B} \in \mathbb{R}^{(p+q) \times (p+q)}$, and $\boldsymbol{\Sigma}_U$ is defined in (14). Let

$$\widetilde{\mathbf{U}}_{n(2)} = \mathbf{U}_{n(2)}\left(\tilde{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right), \quad \widetilde{\Sigma}_{(2)}^{-1} = \Sigma_{(2)}^{-1}\left(\tilde{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right), \tag{20}$$

i.e., the quantities of $\mathbf{U}_{n(2)}$ and $\boldsymbol{\Sigma}_{(2)}$ with $\boldsymbol{\theta}$ substituted by $\tilde{\theta}$. The theorem below summarizes the asymptotic properties of the score statistic.

Theorem 2. Under mild regularity conditions and $H_0 : \boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{(20)}$, the score test statistic $S_n\left(\tilde{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right)$ has an asymptotic χ_q^2 distribution with q degrees of freedom, i.e.,

$$S_n\left(\tilde{\theta}_{(1)}, \boldsymbol{\theta}_{(20)}\right) = n\widetilde{\mathbf{U}}_{n(2)}^{\top}\, \widetilde{\Sigma}_{(2)}^{-1}\, \widetilde{\mathbf{U}}_{n(2)} \rightarrow_d \chi_q^2\,. \tag{21}$$

A sketch of proof is provided in the Supporting Information.

4. Applications

We first investigated the performance of this FRM approach and compared it with the PERMANOVA, then applied it to a study on alcoholic liver disease (ALD). For Monte Carlo (MC) simulations, we set $M = 1,000$ for MC iterations, two-sided type I error rate $\alpha = 0.05$, and sample size (per group) $n_k = 50, 100, 500$ ($k = 1, 2$) for two groups. All analyses were performed within the R software platform (Team, 2017), with code optimized using

Rcpp (Eddelbuettel et al., 2011) for run-time improvement, which is available as Supporting Information.

4.1 Simulation Study

Beta-diversity is a feature summarization for the high-dimensional and zero-inflated counts of taxonomic units extracted from sequence data. Hence, our approach is to first generate those taxonomic abundances, and then compute Beta-diversity distances from the normalized taxonomic abundances. Also, as microbial abundances for each taxonomic unit are usually not independent, common approaches to generate taxonomic abundances from parametric distributions fail to produce life-like microbiome data (Zhang et al., 2017). We thus develop an approach to generate data that resemble real taxonomic abundances based on their empirical cumulative distribution function (eCDF) and copula (See the Supporting Information for details). As this procedure does not involve analytical distributional models, population-level characteristics such as the mean are estimated by Monte Carlo simulation with a large MC size of 5,000.

4.1.1 Simulation Settings.—We generated Beta-diversity outcomes from eCDFs of OTU counts in a study on alcoholic liver disease (Lang et al., 2020b). Chronic alcohol consumption increases intestinal permeability and changes the intestinal microbiota composition, which contributes to the progression of alcohol-related liver disease (ALD). In this study, $n = 85$ subjects including 59 alcoholic hepatitis (AH) patients, 15 alcohol user disorder (AUD) patients, and 11 healthy controls (HC) were enrolled. Fungal ITS sequencing and analysis were conducted using the Illumina MiSeq V3 platform specific for the fungal ITS1 region, resulting in $p = 81$ detected genera. Beta-diversity were computed from the OTU relative abundance vector $\mathbf{Y}_{85\times 81} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{85}]^\top$. For space consideration, we reported results using the Bray-Curtis distance.

Shown in the left-most panel of Figure 1 are eCDFs of Beta-diversity in the three diagnostic groups. The eCDFs are considerably different between the AH and HC as well as AUD and HC group, but less so between the AUD and AH. To illustrate, we combined the AH and AUD patients and simulated OTUs from this combined disease (D) and HC group. Shown in the center of Figure 1 are the eCDFs of observed Beta-diversity for the D and HC group, and in the right-most panel are those of the simulated Beta-diversity for a sample size of $n_k = 500$, which are nearly identical to their original counterparts. The Supporting Figure 1 provides Principal Coordinates Analysis (PCoA) plot, a popular visualizing tool for Beta-diversity (Kruskal and Wish, 1978), which also reveals similar patterns.

To assess whether the data generating procedure retains the important feature of zero-inflated OTUs, we evaluated the average percentage of zero counts in real (93.93%) and simulated OTUs, which are 93.34% ($sd = .004$) for $n_k = 50$; 93.55% ($sd = .003$) for $n_k = 100$; and 94.10% ($sd = .001$) for $n_k = 500$, indicating that the simulated OTUs do reflect the zero-inflated nature of the real OTUs.

4.1.2 Group Comparison.—We first considered group comparisons without any covariate, where the FRM parameterized with an intercept is given by:

$$E\{f(\mathbf{y_i})|\mathbf{x_i}\} = h(\mathbf{x_i}, \boldsymbol{\theta}) = \exp\{\beta_0 + \beta_{22}\delta_{22}(x_\mathbf{i}) + \beta_{12}\delta_{12}(x_\mathbf{i})\}, \\ \mathbf{i} = (i_1, i_2) \in C_2^n, \quad \boldsymbol{\theta} = (\beta_0, \beta_{22}, \beta_{12})^\top, \tag{22}$$

where $n = n_1 + n_2$ with n_k denoting the sample size of group k and $f(\mathbf{y_i}) = d_{i_1, i_2}$ denoting the Beta-diversity outcome for pair $\mathbf{i} = (i_1, i_2) \in C_2^n$. The three types of hypotheses are:

$$\begin{aligned} \text{Within-group:} \;\; & H_{01}: \beta_{22} = 0, \quad \text{vs.} \quad H_{a1}: \beta_{22} \neq 0, \\ \text{Between-group:} \;\; & H_{02}: \beta_{12} = 0, \quad \text{vs.} \quad H_{a2}: \beta_{12} \neq 0, \\ \text{Within- vs. between-group:} \;\; & H_{03}: \beta_{22} = \beta_{12}, \quad \text{vs.} \quad H_{a3}: \beta_{22} \neq \beta_{12}\,. \end{aligned} \tag{23}$$

To assess the performance of the proposed approach for varying sample sizes, we simulated OTUs from a single group based on the eCDF of group D using the copula approach. In this case, all three null hypotheses in (23) hold.

Let $\hat{\boldsymbol{\theta}}^{(m)}$ denote the estimator of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(m)}$ the asymptotic variance from the mth MC iteration, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(asymp)}$ denote the sample mean of $\hat{\boldsymbol{\theta}}^{(m)}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(m)}$, respectively, and let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(emp)}$ denote the sample variance of $\hat{\boldsymbol{\theta}}^{(m)}$. Let $W_n^{(m)}$ denote the Wald statistic in (16) for testing a hypothesis at the mth MC iteration. The type I error rate based on the asymptotic variance is given by $\hat{\alpha}^W = (1/M)\sum_{m=1}^{M} I\left(W_n^{(m)} \geqslant q_{s,0.95}\right)$, where $q_{s,0.95}$ denotes the 95^{th} percentile of a central χ^2 distribution with s degrees of freedom. The score type I error rate $\hat{\alpha}^s$ was computed similarly by replacing $W_n^{(m)}$ with the score statistic in (21) at the mth iteration.

We assess the asymptotic performance by comparing asymptotic and empirical standard errors from $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(asymp)}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(emp)}$ and by comparing $\hat{\alpha}^W(\hat{\alpha}^s)$ and $\alpha = 0.05$.

Shown in Table 1 are estimates (Est.) of $\boldsymbol{\theta}$, asymptotic and empirical standard errors. $\hat{\beta}_{22}$ and $\hat{\beta}_{12}$ were quite close to 0 (true value). The true $\beta_0 = -0.4595$ was obtained by the sample mean of Beta-diversity for a large MC sample size of 5, 000. The estimated $\hat{\beta}_0$'s were close to the truth for all three sample sizes. The asymptotic standard errors were close to their empirical counterparts. As expected, discrepancies became smaller as the sample size increased. But estimates and asymptotic standard errors of $\boldsymbol{\theta}$ were still good for $n_k = 50$.

Shown in Table 2 are type I errors of FRM for the three nulls in (23) and PERMANOVA for the overall group difference. For the FRM, although exhibiting a small upward bias for $n_k = 50$, the Wald type I errors were close to $\alpha = 0.05$ in all three cases. The score tests worked well to reduce bias for $n_k = 50$ and 100 with nearly identical type I errors as the Wald for large sample sizes. PERMANOVA also performed well, albeit with a small downward bias for $n_k = 50$ and 100, which often occurs for small sample sizes (Hemerik et al., 2018).

4.1.3 Group Comparison Accounting for Covariates.

—We illustrate with one continuous and one binary covariate, with the same two diagnostic groups as in (22), the FRM becomes:

$$
\begin{aligned}
&E\{f(\mathbf{y_i})|\mathbf{x_i}, z_\mathbf{i}\} = h(\mathbf{x_i}, z_\mathbf{i}; \boldsymbol{\theta}) = \exp\left(\mathbf{u_i}^\top \boldsymbol{\theta}\right),\\
&\qquad \mathbf{u_i}^\top \boldsymbol{\theta} = \beta_0 + \beta_{22}^d \delta_{22}^d\left(x_\mathbf{i}^d\right) + \beta_{12}^d \delta_{12}^d\left(x_\mathbf{i}^d\right) + \beta_{22}^g \delta_{22}^g\left(x_\mathbf{i}^g\right) + \beta_{12}^g \delta_{12}^g\left(x_\mathbf{i}^g\right)\\
&+ \xi^a g^a\left(z_\mathbf{i}^a\right),\\
&\qquad \boldsymbol{\theta} = \left(\beta_0, \beta_{22}^d, \beta_{12}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top, \quad \mathbf{i} = (i_1, i_2) \in C_2^n,
\end{aligned}
\tag{24}
$$

where $x_\mathbf{i}^d$, $x_\mathbf{i}^g$ and $z_\mathbf{i}^a$ denote the diagnostic group, binary and continuous covariates for each pair $\mathbf{i} \in C_2^n$. In addition to the three null hypotheses comparing diagnostic groups, two new hypotheses can be tested with $H_{04a}: \xi^a = 0$ for the continuous and $H_{04b}: \beta_{22}^g = \beta_{12}^g = 0$ for the binary covariate. Simulation details are provided in the Supporting Information.

Shown in Table 3 are estimates and results for testing the nulls. Again, all estimates were close to their respective true values, and asymptotic standard errors were close to their empirical counterparts. Wald and score type I errors were also close to the nominal value, albeit a bit inflated for the Wald with $n_k = 50$. The gaps between Wald and score type I errors became negligible with large sample sizes.

4.1.4 Power Comparison with the Existing Approach.

—We then compared the power and computational time of the proposed FRM with PERMANOVA to highlight its advantages. Specifically, we compared hypotheses (1) "Between-group" difference with PERMANOVA and (2) "Within-group" difference with 'betadisper' function in *'vegan'* (Oksanen et al., 2013) as a proxy, since PERMANOVA does not directly test this hypothesis. Since it is not straightforward for PERMANOVA to test (3) "Within- vs. Between-group" difference, we did not include this comparison. The simulation details are provided in the Supporting Information. Both permutation-based PERMANOVA and 'betadisper' were conducted with the number of permutations set to 99, 299, 499, and 999, respectively.

Shown in Table 4 are group size, effect size, power, and elapsed time (of one iteration) for comparison. In detecting between-group differences (i.e., location), FRM outperformed PERMANOVA in both power and scalability. Not only did FRM attain much higher power, but it also required far less computing time. For within-group differences (i.e., dispersion), FRM still surpassed 'betadisper' in scalability and achieved slightly higher power. For both PERMANOVA and 'betadisper', the computational time increased dramatically with the increased number of permutations.

4.2 Real Data Analyses

We also applied the proposed FRM to the alcoholic liver disease study (Lang et al., 2020a) to compare Beta-diversity among the original three diagnostic groups. Our goal was to identify the association between the microbiome diversity and diagnostic groups, controlling for demographics. The FRM for diagnostic groups and two covariates of gender and age is:

$$
\begin{aligned}
E\{f(\mathbf{y_i}) \mid \mathbf{x_i}, z_\mathbf{i}\} &= h(\mathbf{x_i}, z_\mathbf{i}; \theta) = \exp\left(\mathbf{u_i}^\top \boldsymbol{\theta}\right), \\
\mathbf{u_i} &= \left(1, \delta_{22}^d\left(x_\mathbf{i}^d\right), \delta_{33}^d\left(x_\mathbf{i}^d\right), \delta_{12}^d\left(x_\mathbf{i}^d\right), \delta_{13}^d\left(x_\mathbf{i}^d\right), \delta_{23}^d\left(x_\mathbf{i}^d\right), \delta_{22}^g\left(x_\mathbf{i}^g\right), \delta_{12}^g\left(x_\mathbf{i}^g\right), g^a\left(z_\mathbf{i}^a\right)\right)^\top, \\
\mathbf{i} &= (i_1, i_2) \in C_2^n, \quad \boldsymbol{\theta} = \left(\beta_0, \beta_{22}^d, \beta_{33}^d, \beta_{12}^d, \beta_{13}^d, \beta_{23}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top,
\end{aligned}
\tag{25}
$$

where β_0 represents the log of mean within-group Beta-diversity for the reference AH group, β_{kk}^d represent the log of mean within-group Beta-diversity differences for AUD ($k = 2$) and HC ($k = 3$) with the AH ($k = 1$), and β_{kl}^d represent the log of mean differences of the respective between-group Beta-diversity of AH and AUD $\left(\beta_{12}^d\right)$, AH and HC $\left(\beta_{13}^d\right)$, AUD and HC $\left(\beta_{23}^d\right)$ compared with the AH, $\beta_{22}^g\left(\beta_{12}^g\right)$ represents the log of mean difference of Beta-diversity comparing female-female (male-female) and the reference male-male pairs, and ξ^a represents the change in the log of mean Beta-diversity per unit increase in age difference (measured by Euclidean distance). Given the relatively small sample sizes for AUD and HC, we report both Wald and score results, as well as Bootstrap results (based on 5, 000 Bootstrap samples) to assess the accuracy of asymptotic results.

The top of Table 5 shows estimates, standard errors (asymptotic under “A. SE” and Bootstrap under “B. SE”), test statistics and p-values (Wald under “W. p”, score under “S. p”, Bootstrap Wald under “B.W. p” and Bootstrap score under “B.S. p”) for the nulls. All Bootstrap standard errors were close to their asymptotic counterparts. For each hypothesis, the test results were consistent, except for a noticeable discrepancy of the score test for β_{33}^d due to the small sample size of HC group ($n_3 = 11$).

AUD had no significant within-group difference in mean diversity compared with the AH ($\hat{\beta}_{22}^d = .226$, p-values range [.419, .662]), but HC had a significantly higher within-group diversity than the AH from Wald test ($\hat{\beta}_{33}^d = .572$, W. p = .002), which is consistent with Figure 1. While the score test for β_{33}^d revealed that more evidence needed to be collected to reject the null (S. p = .130), this discrepancy may be due to the small sample size of HC. However, after Bootstrapping, both Wald and score were consistently significant for β_{33}^d (B.W. p = .007, B.S. p < .0001). All the above results reveal the scientific finding that alcoholic liver disease is associated with reduced microbial diversity. For covariates, age had a positive effect with $\hat{\xi}^a = .006$, both female-female $\left(\hat{\beta}_{22}^g = .125\right)$ and male-female $\left(\hat{\beta}_{12}^g = 0.72\right)$ pairs had higher mean diversity than male-male pairs. None of the covariates were significant.

The bottom of Table 5 includes statistics and p-values. The null of no within-group difference $\left(H_{01}: \beta_{22}^d = \beta_{33}^d = 0\right)$ was rejected consistently by Wald (W. p = .007) and two bootstrap tests (B.W. p = .017, B.S. p < .0001), while the score test was close to being significant with S. p = .071, suggesting a larger sample size may be needed to confirm significance. The null of no between-group difference $\left(H_{02}: \beta_{12}^d = \beta_{13}^d = \beta_{23}^d\right)$ across the three groups was rejected by all tests with the p-values ranging in (.0001, .001].

The between- vs. within-group differences were significant for between-group variability of D-HC and within-group variability of AH-AH pairs: with p-values ranging in (.0001, .006] for $H_{03}^{(2)}: \beta_{13}^d = 0$ (AH-HC vs. AH-AH) and (.0001, .020] for $H_{03}^{(3)}: \beta_{23}^d = 0$ (AUD-HC vs. AH-AH). However, there was no evidence to reject $H_{03}^{(1)}: \beta_{12}^d = 0$ concerning the between-group variability of AUD-AH vs. within-group variability of AH-AH pairs. There was no significant difference across the three gender pair groups (p-values range in [.732, 1]).

The results above were not corrected for multiple comparisons. We also provide FDR corrected results in the Supporting Information by applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the family-wise FDR at 5%, where major conclusions remained unchanged except for $H_{03}^{(3)}: \beta_{23}^d = 0$ (AUD-HC vs. AH-AH), the score test p-value (S. p) was .020 before and .060 after correction.

In summary, both within- and between-group hypotheses detected group differences, driven by the fact that the HC group was rather distinct from the two disease groups. While the within- vs. between-group hypotheses enabled a more comprehensive comparison, the difference between AH-AH and AUD-AUD pairs was not as pronounced, yet any pair involving one subject from HC was significantly different from AH-AH pairs. These specific conclusions underscore the advantages of partitioning the sources of variation under the FRM.

5. Discussion

We developed a new approach to model Beta-diversity utilizing the functional response models (FRM). Unlike conventional approaches such as the PERMANOVA, the proposed FRM can disentangle information carried by Beta-diversity flexibly with the unique interpretations of “mean within-group diversity” for each group and the “mean between-group diversity” between any two groups. This regression approach also provides coefficient estimators for explanatory variables, generating interpretable results on both the direction and size of the effects and leading to more in-depth scientific findings.

In addition, the proposed approach carries far fewer overheads than PERMANOVA in terms of the computational burden. Also, the semiparametric nature of the model enables valid inferences without any parametric assumption on the correlated and non-negative Beta-diversity. Lastly, the approach to simulate life-like OTUs and Beta-diversity allows one to relate simulation study results directly to the performance of the proposed and other statistical models for such data in real studies.

Comparing with other methods for multivariate responses to improve inference of the mean response such as the covariance regression model (Hoff and Niu, 2012), the proposed approach aims to directly model the relationships between Beta-diversity, a complex yet biologically meaningful between-subject attribute, and a set of explanatory variables, which can be within-, between-subject or both, as deemed appropriate by content experts. Also, FRM's ability to control for between-subject confounders, such as a dissimilarity measure comparing subjects' metabolites abundance profile, makes it particularly useful in certain circumstances involving such confounders. Given some recent discussions (Morton et al., 2019) regarding the confounding of sequencing depth, one potential issue in most compositional data analysis is the stochastic nature of sampling reads due to technical variation, yielding a potential confounding effect. If this is the case in some applications, we can alleviate it by modeling Beta-diversity from the absolute abundance (instead of relative abundance) and including the sampling depth as an offset term in the proposed model.

In practice, we suggest conducting both score and Wald tests in applying the proposed model. If the sample size for some groups is relatively small (for example, $n_k < 50$), an additional Bootstrap procedure is recommended. One major limitation of the approach is that it only applies to cross-sectional data. Currently, leveraging semiparametric regression models for longitudinal data, we are working on extending the approach to facilitate analyses of such data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by a DFG fellowship (LA 4286/1–1), an AASLD Clinical and Translational Research Fellowship Award (to S.L.), and NIH grants U01 AA026939, UL1TR001442 and DK120515.

Data Availability Statement

The data that support the findings in this paper are openly available in National Center for Biotechnology Information at https://www.ncbi.nlm.nih.gov/bioproject/, reference number PRJNA517994.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical Society: Series B 57, 289–300.

Dubitzky W,Wolkenhauer O, Yokota H, and Cho K-H (2013). Encyclopedia of systems biology. Springer Publishing Company, Incorporated.

Durack J and Lynch SV (2019). The gut microbiome: Relationships with disease and opportunities for therapy. Journal of Experimental Medicine 216, 20–40. [PubMed: 30322864]

Eddelbuettel D, Francois R, Allaire J, Ushey K, Kou Q, Russel N, et al. (2011). Rcpp: Seamless r and c++ integration. Journal of Statistical Software 40, 1–18.

Hemerik J, Goeman JJ, et al. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. Journal of the Royal Statistical Society Series B 80, 137–155.

Hoff PD and Niu X (2012). A covariance regression model. Statistica Sinica pages 729–753.

Holmes E, Li JV, Athanasiou T, Ashrafian H, and Nicholson JK (2011). Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. Trends in Microbiology 19, 349–359. [PubMed: 21684749]

Kennedy P (2008). A guide to econometrics. John Wiley & Sons.

Kowalski J and Tu XM (2008). Modern applied U-statistics, volume 714. John Wiley & Sons.

Kruskal JB and Wish M (1978). Multidimensional scaling (quantitative applications in the social sciences). Beverly Hills.

Lang S, Duan Y, Liu J, Torralba MG, Kuelbs C, Ventura-Cots, et al. (2020a). Human intestinal mycobiome in alcoholic liver disease targeted loci. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA517994.

Lang S, Duan Y, Liu J, Torralba MG, Kuelbs C, Ventura-Cots, et al. (2020b). Intestinal fungal dysbiosis and systemic immune response to fungi in patients with alcoholic hepatitis. Hepatology 71, 522–538. [PubMed: 31228214]

Legendre P and Gallagher ED (2001). Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280. [PubMed: 28547606]

Lin T, Chen T, Liu J, and Tu XM (2021). Extending the mann-whitney-wilcoxon rank sum test to survey data for comparing mean ranks. Statistics in Medicine 40, 1705–1717. [PubMed: 33398899]

Lozupone C and Knight R (2005). Unifrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71, 8228–8235. [PubMed: 16332807]

McArdle BH and Anderson MJ (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82, 290–297.

Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. (2019). Establishing microbial composition measurement standards with reference frames. Nature Communications 10, 1–11.

National Academies of Sciences, E. and Medicine (2018). Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy. The National Academies Press, Washington, DC.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara R, et al. (2013). Package 'vegan'. Community Ecology Package, Version 2, 1–295.

Roberts DW (2017). Distance, dissimilarity, and mean–variance ratios in ordination. Methods in Ecology and Evolution 8, 1398–1407.

Sørensen TJ (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. I kommission hos E. Munksgaard.

Tang W, He H, and Tu XM (2012). Applied categorical and count data analysis. CRC Press.

Team RC (2017). R core team (2017). r: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria

Wu P, Gunzler D, Lu N, Chen T, Wymen P, and Tu XM (2014). Causal inference for community-based multi-layered intervention study. Statistics in Medicine 33, 3905–3918. [PubMed: 24817513]

Zhang Y, Zhou H, Zhou J, and Sun W (2017). Regression models for multivariate count data. Journal of Computational and Graphical Statistics 26, 1–13. [PubMed: 28348500]

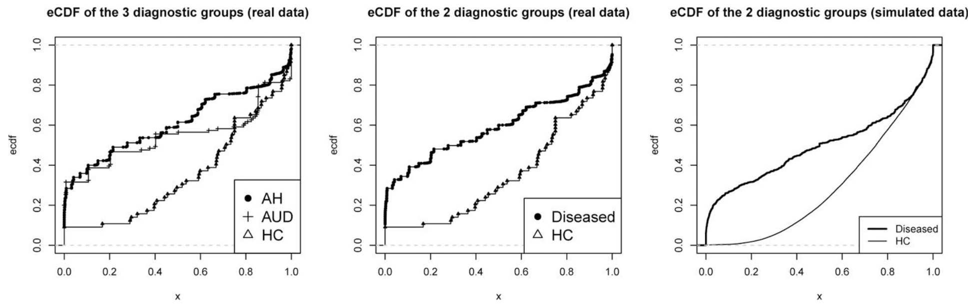

Figure 1.

Empirical cumulative distribution functions (eCDF) of OTU relative abundances for (1) real data of alcoholic hepatitis (AH) patients, alcohol user disorder (AUD) patients, and non-alcoholic controls (HC) (left) (2) real data of combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (middle), and (3) simulated data of combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (right).

Table 1.

MC estimates, standard errors (asymptotic and empirical) for FRM under the null hypotheses, averaged over MC $M = 1,000$ *iterations.*

Under Null Hypotheses			
Parameter	**Est.**	**Std. err**	
		Asymptotic	**Empirical**
	$n_k = 50$		
β_0	−.438	.091	.093
β_{22}	.003	.128	.133
β_{12}	.004	.066	.068
	$n_k = 100$		
β_0	−.452	.066	.065
β_{22}	.0003	.093	.096
β_{12}	.002	.048	.049
	$n_k = 500$		
β_0	−.458	.030	.031
β_{22}	.0007	.043	.043
β_{12}	.0006	.021	.021

Table 2.

Comparison of type I error rates between FRM (based on Wald and Score tests) and PERMANOVA (based on permutation).

	FRM: Type of Hypothesis			PERMANOVA
Sample size n_k	**Within-:** $H_{01} : \beta_{22} = 0$	**Between-** $H_{02} : \beta_{12} = 0$	**Within- vs. Between-** $H_{03} : \beta_{22} = \beta_{12}$	
		Type I Error Rates (Wald)		
50	.045	.081	.087	
100	.046	.063	.071	
500	.047	.053	.057	
		Type I Error Rates (Score)		Type I Error Rates
50	.038	.048	.054	.043
100	.044	.047	.054	.048
500	.047	.051	.053	.051

Table 3.

MC estimates, standard errors (asymptotic and empirical), and type I error rates (Wald and Score) of FRM controlling for covariates under the null hypotheses, averaged over MC $M = 1{,}000$ *iterations.*

Categorical Covariate: Gender (β^g), Continuous Covariate: Age (ξ^a)					
Parameter	**Est.**	**Std. err**		**Type I Error**	
		Asymptotic	**Empirical**	**Wald**	**Score**
$n_k = 50$					
β_0	−.442	.127	.135	.087	.048
β_{22}^d	.003	.130	.139	.059	.055
β_{12}^d	.004	.068	.072	.074	.045
β_{22}^g	.497	.129	.133	.047	.039
β_{12}^g	.501	.066	.069	.084	.056
ξ^a	.500	.098	.097	.050	.037
$n_k = 100$					
β_0	−.456	.085	.083	.057	.046
β_{22}^d	.0005	.094	.097	.060	.055
β_{12}^d	.002	.048	.049	.076	.059
β_{22}^g	.502	.094	.094	.046	.044
β_{12}^g	.502	.048	.048	.064	.046
ξ^a	.500	.056	.055	.048	.044
$n_k = 500$					
β_0	−.456	.039	.041	.057	.056
β_{22}^d	.0003	.043	.044	.050	.050
β_{12}^d	.0004	.022	.022	.049	.046
β_{22}^g	.498	.043	.045	.055	.056
β_{12}^g	.499	.021	.022	.061	.057
ξ^a	.500	.029	.029	.049	.050

Table 4.

Comparisons of power and computational time between FRM and PERMANOVA as well as ‘betadisper’, with the number of permutations set to 99, 299, 499, and 999 for both permutation-based approaches.

“Between-group” difference (location): FRM vs. PERMANOVA											
n k	**Effect Size**	**Power**					**Time for one iteration (*s*)**				
		FRM	**PERMANOVA (#)**				**FRM**	**PERMANOVA (#)**			
			99	**299**	**499**	**999**		**99**	**299**	**499**	**999**
50	.322	.637	.152	.168	.172	.176	.009	.017	.051	.079	.180
100	.346	.905	.383	.423	.431	.441	.024	.078	.238	.408	.878
200	.346	.994	.892	.927	.922	.921	.108	.332	1.051	1.929	3.642
“Within-group” difference (dispersion): FRM vs. ‘betadisper’											
n k	**Effect Size**	**Power**					**Time for one iteration (*s*)**				
		FRM	**Betadisper (#)**				**FRM**	**Betadisper (#)**			
			99	**299**	**499**	**999**		**99**	**299**	**499**	**999**
50	.352	.698	.662	.698	.697	.691	.009	.015	.040	.062	.121
100	.366	.956	.914	.922	.928	.925	.024	.015	.041	.064	.126
200	.362	1.000	.996	1.000	.999	.998	.108	.020	.049	.075	.153

Table 5.

Estimates, asymptotic standard errors (A. SE), Bootstrap standard errors (B. SE) based on $B = 5{,}000$ *Bootstrap samples, Wald statistics, Score statistics, Wald p-values (W. p), Score p-values (S. p), Bootstrap Wald p-values (B.W. p) and Bootstrap Score p-values (B.S. p) for the real study data using FRM, including covariates.*

Categorical Covariate: Gender (β^g), Continuous Covariate: Age (ξ^a)

Parameter	Est.	Std. err		Statistic		p-value			
		A. SE	B. SE	Wald	Score	W. p	S. p	B.W. p	B.S. p
β_0	−1.042	.215	.226	23.485	13.630	<.0001	.0002	<.0001	<.0001
β_{22}^d	.226	.302	.290	.560	.442	.454	.506	.419	.662
β_{33}^d	.572	.186	.201	.416	2.294	.002	.130	.007	<.0001
β_{12}^d	.114	.193	.174	.350	.331	.554	.565	.519	.674
β_{13}^d	.634	.173	.183	13.409	7.456	<.0001	.006	.002	<.0001
β_{23}^d	.672	.180	.190	14.002	5.408	<.0001	.020	.0004	<.0001
β_{22}^g	.125	.189	.175	.436	.399	.509	.528	.477	.613
β_{12}^g	.072	.121	.111	.357	.356	.550	.551	.511	.583
ξ^a	.006	.005	.005	1.723	1.479	.189	.224	.184	.348

Hypothesis		Statistic		p-value			
		Wald	Score	W. p	S. p	B.W. p	B.S. p
Within-	$H_{01}: \beta_{22}^d = \beta_{33}^d = 0$	9.865	5.295	.007	.071	.017	<.0001
Between-	$H_{02}: \beta_{12}^d = \beta_{13}^d = \beta_{23}^d$	19.009	28.477	<.0001	<.0001	.001	<.0001
Within- vs. Between-	$H_{03}^{(1)}: \beta_{12}^d = 0$	.350	.331	.554	.565	.519	.674
	$H_{03}^{(2)}: \beta_{13}^d = 0$	13.409	7.456	<.0001	.006	.002	<.0001
	$H_{03}^{(3)}: \beta_{23}^d = 0$	14.002	5.408	<.0001	.020	.0004	<.0001
Covariates	$H_{04a}: \xi^a = 0$	1.723	1.479	.189	.224	.184	.613
	$H_{04b}^{(1)}: \beta_{22}^g = 0$	.436	.399	.509	.528	.477	.583
	$H_{04b}^{(2)}: \beta_{12}^g = 0$	.357	.356	.550	.551	.511	.348
	$H_{04b}: \beta_{22}^g = \beta_{12}^g = 0$	.621	.241	.733	.886	.732	1.000