Table 1: Comparison of different mediation analysis methods for microbiome data. Abbreviations: SEM = Structural Equation Model; ilr = isometric logratio transformation; alr = additive logratio transformation; LRT = log-ratio transformation; LM = linear regression model; Taxa = using raw taxonomic units directly as mediator without applying compositional data transformations. Pseudocount refers to the technique of adding a small constant (usually 0.5) to zero counts to avoid mathematical issues in log-ratio transformations or other compositional analyses.

	MedTest	**MODIMA**	**CCMM**	**SparseMCMM**	**IsometricLRTMM**	**microHIMA**
Framework	Distance-based	Distance-based	Counterfactual	Counterfactual	SEM	SEM
Mediator	PCoA	Distance matrix	alr	alr	ilr	ilr
Overall ME	✓	✓	✓	✓	×	×
Mediator-wise ME	×	×	✓	×	✓	✓
Zero counts	-	-	Pseudocount	Pseudocount	Pseudocount	Pseudocount
Regularization	×	×	×	L1 penalty	de-biased Lasso	de-biased Lasso
Model for mediator	LM	LM	Compositional algebra	Dirichlet regression	LM	Log-linear regression
Model for response	LM	LM	Linear log-contrast	Linear log-contrast	LM	LM

	MedZIM	**MicroBVS**	**PhyloMed**	**LDM**	**PERMANOVA-med**	**NPEM**
Framework	Counterfactual	Counterfactual	Phylogenetic tree	Inverse regression	Distance-based	Information-theoretic
Mediator	Taxa	Taxa	LRT	Taxa	Distance matrix	Taxa
Overall ME	×	✓	×	✓	✓	×
Mediator-wise ME	✓	✓	✓	✓	×	✓
Zero counts	-	Pseudocount	Pseudocount	-	-	-
Regularization	×	Bayesian priors	×	✓	×	×
Model for mediator	LM	Log-linear regression	Log-linear regression	Inverse regression	Distance-based	Information-based
Model for response	LM	LM	Log-linear regression	Inverse regression	Distance-based	Information-based

Review of Mediation Analysis on Microbiome

Haoyi Zheng, Ziman Jiang

May 2024

1 Notation

- T: Treatment
- M: Microbiome, p-dimension
- N: Total counts
- Y: Outcome
- X: Covariates
- n: Sample Size

2 Introduction

3 method

Performing mediation analysis on microbiome data presents several unique challenges compared to omics studies. The high-dimensional and compositional nature of microbiome mediators, along with their sparsity and zero-inflated characteristics, imposes significant challenges on mediation models. Traditional mediation analysis methods are often constrained by these features of microbiome data, limiting the power of these methods. The following section introduces some methodological developments designed to address these research challenges.

3.1 Distance-Based Dimensionality Reduction Methods

In this section, we introduce three methods designed to address the challenges of mediation analysis in microbiome data. The methods are Distance-Based Omnibus Test of Mediation Effect (MedTest), Multivariate Omnibus Distance Mediation Analysis (MODIMA), and Nonparametric Entropy Mediation (NPEM). We will start with the commonalities of these three approaches and then introduce them individually.

These three methods are all SEM-based methods. A typical structural equation model (SEM) consists of a measurement model and a structural model. In Mediation Analysis, SEM is able to effectively separate direct and indirect effects, capture the effects of potential mediating variables through measurement modeling, and identify the relationships between independent variables, dependent variables, and mediating variables through structural modeling. The SEM mediation framework for assessing mediation effects here is primarily based on the findings of Baron and Kenny using the product of coefficients approach.

Baron and Kenny's framework for mediator analysis is based on the Single Mediator Model (SMM), a system of three variables describing the relationship between exposure T, response Y, and mediator M. SMM assumes these relationships can be test by linear regression. Let M be an $n \times m$ count matrix representing the abundances of m OTUs in n samples, T be an $n \times 1$ vector of the independent variable, and Y be an $n \times 1$ vector of the outcome variable. The method sets up the microbiome mediates the effect of T on Y through some unknown microbiome feature vector $f^{(l)}(M)$. (missing explanation)Microbiome features can be either the abundance or prevalence of a taxon or a weighted average of several functionally related OTUs. In practice, it is difficult to obtain. The mediation model can be written as:

$$Y = T\gamma' + \epsilon$$

$$f^{(l)}(M) = T\alpha_l + \epsilon_l' \quad (l = 1, \ldots, L)$$

$$Y = \sum_{l=1}^{L} f^{(l)}(M)\beta_l + T\gamma + \epsilon''$$

γ',γ represent the total effect and the direct effect of the independent variable T on the outcome Y.

Here the hypothesis test for the mediating effect is to test the product of the coefficients. The null hypothesis H_0 states that there is no mediation effect, meaning the mediator variable M does not explain the relationship between the independent variable T and the outcome variable Y. It is expressed as: (for flm question? global test or individual test ;talk about the two method)

$$H_0 : \alpha_l\beta_l = 0 \text{ for } \forall f^{(l)}(M)$$

This implies that for all microbiome feature vectors $f^{(l)}(M)$, the product of the coefficients $\alpha_l\beta_l$ is zero(not interpretation, either T or other feature not associated). In other words, the effect of T on Y is not mediated by M. And the alternative Hypothesis H_1 is that for at least one microbiome feature vector $f^{(l)}(M)$, the product of the coefficients $\alpha_l\beta_l$ is not zero. (at least two pathway exists)

Furthermore, The first two methods chose the Permutation test for Mediation Effect Testing. The permutation test for mediation effects is a widely used nonparametric method that typically evaluates the significance of mediation by iteratively reshuffling the independent variables and recalculating the mediation effects. Permutation tests provide control over Type I error rates, especially in small samples. And do not rely on the distribution assumptions, allowing them to be generally more robust to outliers and deviations from the typical data structure than parametric tests. The following is a conventional permutation test step for the mediation effects:

1. Define the test statistic $S = \alpha_l\beta_l$.(without hat not an test)
2. Compute the observed test statistic $S = \hat{\alpha}_l\hat{\beta}_l$. (otherwise test one,S need sub, for everypair do the sum of squared)
3. For $j = 1, 2, \ldots, B$:
   (a) Permute the values of T to generate permuted datasets $T^{(j)}$.
   (b) Recompute the test statistic $S^{(j)} = \hat{\alpha}_l^{(j)}\hat{\beta}_l^{(j)}$.
4. Compare the observed statistic with the distribution of permuted statistics:

$$p = \frac{1}{B}\sum_{j=1}^{B} 1(S \leq S^{(j)})$$

   where $1(\cdot)$ is the indicator function.

3.1.1 MedTest, PCoA mediation analysis for microbiome

Medtest is a distance-based nonparametric approach to test the mediation effects proposed by Zhang et al (2018). By using sample-wise distance matrices and principal analysis, this method effectively reduces the dimension of the microbiome mediating variables, instead of working with the original OTU data. In addition the use of multiple distance measures by MedTest (both phylogenetic tree-based distances and non-phylogenetic tree-based distances) allows for better identification of specific mediating effects, which improves the robustness and power of the test.

In the paper by Zhang et al., the microbiome feature vector $f^{(l)}_{M\,(n\times 1)}$ is defined as a scalar function that maps from the high-dimensional OTU abundance space $\mathbb{R}^m$ to the real number space $\mathbb{R}$

$$f_M^{(l)} : \mathbb{R}^m \to \mathbb{R}$$

where n is the number of samples and m is the number of OTUs. Specifically, $f_M^{(l)}$ represents the l-th feature vector extracted from the distance matrix of the OTU using feature extraction methods such as Principal Coordinate Analysis (PCoA). The purpose of using $f_M^{(l)}$ is to effectively detect mediation effect between the exposure variable T and the outcome variable Y by reducing dimension and capturing

non-linear relationships in high-dimensional microbiome data. Below is a detailed explanation of the definition of $f_M^{(l)}$.

Let OTU abundance data matrix $M_{(n\times m)}$ is an $n \times m$ count matrix. The distance matrix $D_{(n\times n)}$ is computed between samples, where each element d_{ij} represents the distance between the i-th and j-th samples. Next, define the Double-Centered Matrix $G_{(n\times n)}$ as follows:

$$G = \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\cdot\right) A \cdot \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)$$

where $A_{(n\times n)} = (a_{ij}) = \left(-\frac{1}{2}d_{ij}^2\right)$, I is the identity matrix and $\mathbf{1}$ is a vector of 1's. The eigenvectors $(u_1, u_2, \ldots, u_L)$ and eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_L)$ are obtained by eigenvalue decomposition of the double-centered matrix G. Here, L represents the number of eigenvectors associated with positive eigenvalues, which are used as the microbiome features $f_{M\,(n\times L)}$.

The mediation model in the paper can be summarized as follows:

$$\begin{aligned} Y &= \boldsymbol{\beta} f_M + \gamma T + \epsilon'' \\ f_M &= \boldsymbol{\alpha} T + \epsilon' \end{aligned}$$

where $\boldsymbol{\beta}$ represents the effect of the microbiome feature f_M on the outcome Y, γ represents the direct effect of the exposure variable T on the outcome Y, $\boldsymbol{\alpha}$ represents the effect of the exposure variable T on the microbiome feature f_M and $\epsilon, \epsilon', \epsilon''$ are error terms. In this model, the mediation effect is the effect of T transmitted through M on Y, defined as the product of the path coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The null hypothesis H_0 indicates that there is no mediation effect through any microbiome feature f_M, which is $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} = 0$. This means that the independent variable T does not affect the outcome Y through the microbiome feature f_M. The alternative hypothesis H_1 suggests a significant mediation effect, which is $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} \neq 0$.

The test statistic S is used to evaluate the significance of the mediating effect as defined below:

$$S = \sum_{l=1}^{L} \lambda_l \left|\hat{\alpha}_l \hat{\beta}_l\right|$$

The coefficients $\hat{\alpha}_l$ and $\hat{\beta}_l$ are calculated by regression model, where $\hat{\alpha}_l$ represents the effect of the exposure variable T on the l-th feature vector $f_M^{(l)}$, and $\hat{\beta}_l$ represents the effect of the l-th feature vector $f_M^{(l)}$ on the outcome variable Y, and λ_l are the corresponding eigenvalues reflecting the importance of each feature vector in the total variation.

The whole Medtest framework considers different distance measures to detect mediation effects. The reason is that the use of multiple distance measures can capture a variety of characteristics and ecological differences in microbiome data, thus providing a more comprehensive and detailed analysis.

The method first computes distance matrices between samples using various distance metrics (such as weighted and unweighted UniFrac distances, Jaccard distance, and Bray-Curtis distance). Then, double-centering and eigenvalue decomposition are performed for each distance matrix and test statistics are computed for each distance metric. Finally, the results of multiple distance measures are combined using the minimum p-value method, and Bonferroni correction is performed to control the Type I error rate.

This method effectively handles high-dimensional, sparse, and non-normal microbiome data, making it suitable for complex microbiome mediation analysis. However, MedTest requires high-quality data, and its power may critically depend on whether the exposure-microbiome association and the microbiome-outcome association coincide at the same set of principal components. When the true mediators in the community are rare taxa, the principal components may not effectively capture the variation at these mediators. In addition the biological meanings of the feature vectors $f_M^{(l)}$ obtained by this distance-based method are ambiguous poorly interpretable compared to models that use the raw data directly. Moreover distance-based method usually can only detect the overall mediation effect, but not the mediation effect of each feature vector separately.

3.1.2 MODIMA, a Method for Multivariate Omnibus Distance Mediation Analysis

MODIMA is another distance-based framework designed to test the effects of multivariate distance mediation allowing for multivariate exposures, responses, and mediators. Firstly, MODIMA utilizes pairwise distance matrices to reduce the dimension of high-dimensional variables, and then uses distance

correlation and partial distance correlation to represent the exposure-mediator-response relationship, thus MODIMA is capable of capturing the relationship between variables even in omics data with many limitations, such as microbiome data, which are high-dimensional, compositional, and over-dispersed.

The framework of MODIMA is primarily based on the Mediation Analysis method published by Boca et al[**?**]. In the article published by Boca et al, the null hypothesis of mediation analysis is stated in a less common form as $H_0 : \rho_{TM}\rho_{MY|T} = 0$,where ρ_{TM} is the Pearson Correlation between T and Y which corresponds to α and $\rho_{MY|T}$ is the Conditional Pearson Correlation between M and Y which corresponds to β. The null hypothesis implies that at least one of the paths in the mediation analysis has no effect. Specifically, it means that either the relationship between the exposure T and the mediator M (ρ_{TM}), or the relationship between the mediator M and the outcome Y controlling for the exposure T ($\rho_{MY|T}$), is absent. Rejection of this hypothesis suggests the existence of a mediating effect, where the exposure influences the outcome through the mediator.

The whole test statistic of MODIMA is obtained by first calculating the corresponding pairwise distance matrices $d(\cdot)$ of the variables, and all hypothesis testing and inference is based on it. Depending on the nature of these variables, the distance metric may be different for each variable. The common distance matrices include the Euclidean distance, which is the distance method selected in the simulation part of the article.

Next, inspired by Székely and Rizzo's series of energy statistics-based nonparametric tests of covariance and correlation, MODIMA uses distance correlation ($dCor$) and partial distance correlation ($pdCor$) [**?**] instead of Pearson correlation to capture the relationship between multivariate random variables. Distance Correlation ($dCor$) is used to measure the correlation between the distance matrices $d(T)$ and $d(Y)$. It is defined as follows:

$$dCor(d(T), d(Y)) = \begin{cases} \frac{V(d(T),d(Y))}{\sqrt{V(d(T),d(T))\cdot V(d(Y),d(Y))}}, & \text{if } V(d(T),d(T))\cdot V(d(Y),d(Y)) > 0; \\ 0, & \text{if } V(d(T),d(T))\cdot V(d(Y),d(Y)) = 0. \end{cases}$$

Here, $V(d(T), d(Y))$ is the distance covariance, which measures the average squared difference between all pairwise distances of $d(T)$ and $d(Y)$. The partial distance correlation ($pdCor$), which measures the correlation in distance matrices $d(M)$ and $d(Y)$, controlling for $d(T)$, is described as follows:

$$pdCor(d(M), d(Y); d(T)) = \frac{dCor(d(M), d(Y)) - dCor(d(M), d(T)) \cdot dCor(d(Y), d(T))}{\sqrt{1 - (dCor(d(M), d(T)))^2}\sqrt{1 - (dCor(d(Y), d(T)))^2}}$$

The partial distance correlation ($pdCor$), which measures the correlation in vectors M and Y, controlling for T, is described below:

$$\text{pdCor}(M, Y; T) = \frac{(P_{T^\perp}(M) \cdot P_{T^\perp}(Y))}{|P_{T^\perp}(M)||P_{T^\perp}(Y)|}$$

The MODIMA test statistic can be obtained by multiplying these two quantities as follows:

$$S_d(d(T), d(M), d(Y)) = dCor(d(T), d(M)) \cdot pdCor(d(M), d(Y); d(T)) \tag{3.1}$$

The permutation testing approach for the MODIMA method is also based on the work of Boca et al[**?**]. It obtains the empirical distribution of the test statistic by destabilizing the relationship between the exposure and the mediator or the relationship between the mediator and the outcome. If the relationship between exposure and mediator is weak, the rows and columns of the exposure distance matrix $d_X(X)$ will be displaced. If the relationship between the mediator and the outcome is strong, the rows and columns of the response distance matrix $d_Y(Y)$ are permuted.

Compared to MedTest, MODIMA allows the computation of multivariate exposures and outcomes by calculating the distance matrices. It utilizes partial distance correlation in energy statistics to better capture nonlinear relationships, instead of traditional SEM methods that typically capture linear effects, thus providing higher statistical power in high-dimensional environments. However, the limitation of MODIMA compared to MedTest is that it only applies to specific distance metrics, rather than pooling the results of the analysis of multiple metrics. In addition, based on the simulation results, it is apparent that the choice of the distance metric affects the performance of the model, and MODIMA is unable to select the appropriate distance based on the different dataset. If a distance metric is chosen that is not suitable for a particular data structure or relationship, the analysis results will be affected.

3.2 compositional

3.2.1 CCMM

CCMM (Sohn and Li 2019) is the Causal Compositional Mediation Model, which is used to estimate direct and indirect (or Mediation) effects even if the mediator has a compositional and high - dimensional nature. The novelty of the approach is the use of a counterfactual framework for defining and assessing mediation effects, and the handling of compositional data from microbiome and metagenomic studies. The description of this approach first focuses on how to derive the overall framework of the mediation model using a counterfactual framework, and then the hypothesis test of the mediation effect and the derivation of the parameters.

The counterfactual framework, also known as the potential outcome framework, involves defining potential outcomes for each unit under different treatments. For unit i, let T_i denotes a treatment or exposure variable, $M_{i(k\times 1)}$ be a vector of k compositional mediators, Y_i be an outcome. Let $M_i(t)$ be the potential mediator under $T_i = t$, and $Y_i(t, m)$ the potential outcome under $T_i = t$ and $M_i = m$. In addition denote the observed variables as $M_i = M_i(T_i)$ and $Y_i = Y_i(T_i, M_i(T_i))$. Based on these definitions, the causal direct effect $\zeta(\tau)$ and the causal total indirect effect $\delta(\tau)$ can be defined to capture the effects of treatment on outcome through different pathways, as follows

$$\begin{aligned}\zeta(\tau) &= \mathbb{E}[Y_i(t, M_i(\tau)) - Y_i(t_0, M_i(\tau)) \mid X_i = x] \\ \delta(\tau) &= \mathbb{E}[Y_i(\tau, M_i(t)) - Y_i(\tau, M_i(t_0)) \mid X_i = x]\end{aligned} \tag{3.2}$$

Here, t and t_0 represent different treatment conditions, τ denotes a specific treatment value, and X_i denotes covariates. The Causal Direct Effect (CDE) captures the impact of changing the treatment from $T = t_0$ to $T = t$ on the outcome, while keeping the mediator M constant. In contrast, the Causal Indirect Effect (CIE) measures the effect of the mediator's change from $M_i(t_0)$ to $M_i(t)$ on the outcome, with the treatment T held constant.

Before introducing the compositional mediation model, two transformations should be defined. For two compositions $\eta, \zeta \in \mathcal{S}^{k-1}$, the perturbation operator is defined by:

$$\eta \oplus \zeta = \left(\frac{\eta_1\zeta_1}{\sum_{j=1}^k \eta_j\zeta_j}, \frac{\eta_2\zeta_2}{\sum_{j=1}^k \eta_j\zeta_j}, \ldots, \frac{\eta_k\zeta_k}{\sum_{j=1}^k \eta_j\zeta_j} \right)^T,$$

and the power transformation for a composition η by a scalar ν is given by:

$$\eta^\nu = \left(\frac{\eta_1^\nu}{\sum_{j=1}^k \eta_j^\nu}, \frac{\eta_2^\nu}{\sum_{j=1}^k \eta_j^\nu}, \ldots, \frac{\eta_k^\nu}{\sum_{j=1}^k \eta_j^\nu} \right)^T.$$

This power transformation helps in capturing the effect of the treatment on the mediator within the compositional framework. With these transformations, the compositional mediation model is defined as follows:

$$M_i = (m_0 \oplus a^{T_i} \oplus h_1^{X_{i1}} \oplus \cdots \oplus h_q^{X_{iq}}) \oplus U_{1i} \tag{3.3}$$

$$Y_i = c_0 + cT_i + (\log M_i)^T b + X_i^T g + U_{2i} \tag{3.4}$$

The equation (3.3) can be understood more intuitively by applying an additive logratio transformation. For vector $\boldsymbol{\eta}$ with compositional property, define $alt(\cdot)$ as follows:

$$\mathrm{alt}(\boldsymbol{\eta}) = \left(\log\frac{\eta_1}{\eta_k}, \log\frac{\eta_2}{\eta_k}, \ldots, \log\frac{\eta_{k-1}}{\eta_k} \right)^\top.$$

By applying the $alt(\cdot)$ on both sides of equation (3.3), the result is

$$\mathrm{alt}(M_i) = \mathrm{alt}(m_0) + T_i\mathrm{alt}(a) + \sum_{r=1}^{q} X_{ir}\mathrm{alt}(h_r) + \mathrm{alt}(U_{1i}).$$

So the variables in equation(3.3) can be interpreted as: m_0 and c_0 are the baseline mediator and outcome level; a, b, c are path coefficients; $h_1, \ldots, h_q$ and g are nuisance parameters corresponding to X_i; a^{T_i} represents the treatment effect, $h_1^{X_{i1}}, \ldots, h_q^{X_{iq}}$ are the effects of covariates $X_{i1}, \ldots, X_{iq}$, and U_{1i} and U_{2i} is the error term.

Equations 3.3 and 3.4 describe, respectively, the effect of treatment T_i on the mediator M_i, and the combined effects of the mediator M_i and treatment T_i on the outcome Y_i. This compositional model allows for understanding the direct and indirect effects of treatment on outcomes while controlling for other covariates. And the use of perturbation operators and power transformations is crucial to deal with the compositional nature of the data. Within this compositional model, the causal direct effects (CDE) and causal indirect effects (CIE) can be redefined as follows:

$$\begin{aligned} \text{CDE} &\equiv \mathbb{E}[Y_i(t, \log M_i(\tau)) - Y_i(t_0, \log M_i(\tau))|X_i = x] \\ &= c(t - t_0) \\ \text{CIE} &\equiv \mathbb{E}[Y_i(\tau, \log M_i(t)) - Y_i(\tau, \log M_i(t_0))|X_i = x] \\ &= (\log a)^T b(t - t_0) \end{aligned} \tag{3.5}$$

By definition, the causal total indirect effect $\delta(\tau)$ represents the effect of changes in the mediator M_i on the outcome Y_i. From the formula for $\delta(\tau)$, we can see that if $\delta(\tau)$ is zero, then $(\log a)^T b(t - t_0)$ must also be zero.(repeat) In this case, if $(t - t_0)$ is non-zero, then $(\log a)^T b$ must be zero. Therefore, the null hypothesis H_0 of total compositional mediation effect can be expressed as:

$$H_0 : (\log a)^T b = 0 \tag{3.6}$$

and the null hypothesis of component-wise mediation effect can be expressed as:

$$H_0 : \log(ka_j)b_j = 0 \quad \forall j \in \{1, 2, \ldots, k\} \tag{3.7}$$

Failing to reject the null hypothesis (3.6) implies that the total compositional mediation effect is not significant, indicating that the mediator M_i does not significantly transmit the effect of treatment T_i to the outcome Y_i. Rejecting it suggests a significant indirect path through M_i. Similarly, the null hypothesis (3.7) tests whether any individual mediator component M_{ij} significantly affects Y_i. Failing to reject it implies none of the individual components are significant mediators; rejecting it for any component j indicates that specific M_{ij} has a significant mediation effect.

To test the null hypothesis (3.6) and (3.7), the Sobel test can be employed. The Sobel test for the total compositional mediation effect involves estimating the product of the coefficients $\log \hat{a}$ and $\hat{b}$ and their corresponding variances. The Sobel test statistic is given by:

$$S = \frac{(\log \hat{a})^T \hat{b}}{\sqrt{\hat{b}^2 \sigma^2_{\log \hat{a}} + (\log \hat{a})^2 \sigma^2_{\hat{b}}}}$$

where $\sigma_{\log \hat{a}}$, $\sigma_{\hat{b}}$ are the variance for the estimates of $\log \hat{a}$ and $\hat{b}$, respectively. The article provides methods for estimating these variables using optimization techniques.

The method CCMM provides a robust algebraic proof by testing the mediating effect through a counterfactual framework. Also, the mediators are handled by logratio transformation according to the nature of microbiome data, which removes the inherent constraints of the data to make it suitable for regular statistical analyses. In addition, CCMM can not only test the total mediation effect but also for each mediator. CCMM also has limitations, the counterfactual framework needs to satisfy extra assumptions. The existence of multiple null values in the microbiome data can have an impact on the data transformation process. These are issues that CCMM does not address.

3.2.2 MedZIM

To address the issue of mediation analysis under zero-inflated data structures, Z Li et al developed a novel mediation analysis method called MedZIM. This method is based on the potential outcomes framework and is designed for zero-inflated distributions. In this approach, the mediator variable is assumed to have a zero-inflated distribution (e.g., ZIB distribution) and the mediation effect can be decomposed into two components: one component is the mediation effect caused by the change in numerical value, that is, the difference in abundance level in the presence of the mediating variable. The other component is the effect caused by a change in the mediator variable from a zero to a non-zero state, i.e. the presence or absence of microbial taxa.

To construct the mediation model with zero-inflated mediator variables, first assume that there is a continuous outcome variable Y, a mediator variable M representing the relative abundance (RA) of

microbial taxonomic units, and an independent variable T. The outcome variable Y depends on the mediator M and the independent variable T through the following regression equation:

$$Y = \beta_0 + \beta_1 M + \beta_2 1(M > 0) + \beta_3 T + \beta_4 T 1(M > 0) + \beta_5 TM + \epsilon \tag{3.8}$$

where $1(\cdot)$ is an indicator function, the random error ϵ follows a normal distribution $N(0, \delta)$, and $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are regression coefficients. Then there is also need to model the correlation between M and X. For the mediator M that satisfy a zero-inflated beta (ZIB) distribution, the distribution can be written as a mixed distribution with two-part density function as:

$$f(M = m; \theta) = \begin{cases} \Delta, & m = 0 \\ (1 - \Delta) \cdot \frac{m^{\mu\phi - 1}(1-m)^{(1-\mu)\phi - 1}}{B(\mu\phi, (1-\mu)\phi)}, & m > 0 \end{cases}$$

where $\theta = (\mu, \phi, \Delta)^T$, Δ represents the probability of M taking the value 0, $B(\cdot, \cdot)$ is the beta function, μ and ϕ are the mean and dispersion parameters of the Beta distribution of the nonzero part of M, respectively. To describe the relationship between the parameters θ of the mediator M and the independent variable X, assume that the parameters θ depend on X through the following equation:

$$T(\theta) = \nu_0 + \nu_1 X \tag{3.9}$$

where $T : \mathbb{R}^K \to \mathbb{R}^K$ is a known one-to-one transformation of the parameter vector θ, ν_0 and ν_1 are K-dimensional parameter vectors. For the ZIB distribution, the parameter transformations are as follows: The zero-inflation parameter Δ and mean parameter μ are transformed by a logit function which depends on X, the dispersion parameter ϕ is assumed to have a constant value independent of X. Equations (3.8) and (3.9) together form the full mediation model of this paper.

Under the potential outcomes framework, the article defines the natural indirect effect (NIE), which is mediating effect. The total effect of the exposure variable X is equal to the summation of NIE and NDE. Let M_x denote the value of M if X equals x. Let Y_{xm} denote the value of Y if $(X, M) = (x, m)$. The average NIE for X changing from x_1 to x_2 is defined as:

$$\text{NIE} = E(Y_{x2M_{x2}} - Y_{x2M_{x1}})$$

By plugging the equations (3.8) and (3.9) into the above definitions, we can obtain the following formulas:

$$\begin{aligned} \text{NIE} &= E(Y_{x2M_{x2}}) - E(Y_{x2M_{x1}}) \\ &= (\beta_1 + \beta_5 x2)(E(M_{x2}) - E(M_{x1})) + (\beta_2 + \beta_4 x2)(E(1(M_{x2} > 0)) - E(1(M_{x1} > 0))) \\ &= \text{NIE}_1 + \text{NIE}_2 \end{aligned}$$

Where NIE_1 can be interpreted as the mediation effect caused by the change of the mediator on its numerical value and NIE_2 can be interpreted as the mediation effect due to change of the mediator from a zero state to a non-zero state.

In this study, maximum likelihood estimation (MLE) was used for parameter estimation. First, the data were divided into two groups (non-zero mediated effects group and zero mediated effects group); then, the log-likelihood contribution of each group was calculated; and finally, the log-likelihood function was maximized to obtain parameter estimates. These estimates were used to calculate the natural indirect effect (NIE) and to obtain confidence intervals by delta method or bootstrapping.

The MedZIM method allows for a decomposition of the mediating effect into two parts: one attributable to the change in positive relative abundance, and the other attributable to the binary change from zero to non-zero states. This decomposition allows for a more detailed analysis of zero-inflated data.

3.2.3 microHIMA

In contrast to methods primarily used to test for overall mediation effects, Zhang et al proposed a statistical procedure for selecting individual taxa that mediate the exposure-outcome relationship by using isometric log-ratio transformations.

Assuming that the microbiome in each sample consists of d taxa, its relative abundances are denoted by the vector $M = (M_1, \ldots, M_d)'$. The relative abundance M lies in a specific mathematical space called

the 'simplex' space. This 'simplex' space is characterized by the fact that all components (i.e., relative abundances) are positive and they sum to one. The simplex space is defined as follows:

$$\mathcal{S}_d = \left\{ x = (x_1, \ldots, x_d)' : \sum_{k=1}^{d} x_k = 1; x_k > 0, k = 1, \ldots, d \right\}$$

Due to the compositional property of relative abundance, many statistical models in Euclidean space are not applicable. To solve this problem, the authors used the isometric log-ratio (ilr) transformation to convert compositional data from simplex space $\mathcal{S}_d$ to Euclidean space R^{d-1}, and the ilr transformations for M are defined as follows:

$$\tilde{M}_k = \sqrt{\frac{d-k}{d-k+1}} \ln \frac{M_k}{\sqrt[d-k]{\prod_{j=k+1}^{d} M_j}}, \quad k = 1, \ldots, d-1$$

Each transformed variable $\tilde{M}_k$ is calculated from the ratio of M_k to the geometric mean of the remaining components. This allows the definition of a high-dimensional linear mediation model in Euclidean space:

$$\begin{aligned} E(Y|X, \mathbf{Z}, \tilde{\mathbf{M}}) &= \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{d-1} \tilde{M}_{d-1} + \mathbf{Z}'\theta, \\ E(\tilde{M}_k|X, \mathbf{Z}) &= \alpha_k X + \mathbf{Z}'\eta_k, \quad k = 1, \ldots, d-1 \end{aligned} \tag{3.10}$$

where $\tilde{\mathbf{M}} = (\tilde{M}_1, \ldots, \tilde{M}_{d-1})'$, X is the exposure variable, $\mathbf{Z} = (Z_1, \ldots, Z_q)'$ is a vector of covariates; γ represents the direct effect of X on Y adjusting for $\tilde{\mathbf{M}}$ and $\mathbf{Z}$; α_k represents the relation between X and the mediator $\tilde{M}_k$ adjusting for $\mathbf{Z}$; β_k represents the relation between $\tilde{M}_k$ and Y adjusting for the effects of X, $\mathbf{Z}$, and other mediators; $\theta = (\theta_1, \ldots, \theta_q)'$ and $\eta_k = (\eta_{k1}, \ldots, \eta_{kq})'$ are regression coefficients for $\mathbf{Z}$. The product $\alpha_k \beta_k$ represents the *indirect effect* of X on Y, which is transmitted through the mediator $\tilde{M}_k$.

Due to the nature of the ilr transformation, $\tilde{M}_2, \ldots, \tilde{M}_{d-1}$ are variables that are generated after removing the information of M_1. This means that $\tilde{M}_2, \ldots, \tilde{M}_{d-1}$ do not directly relate to M_1 but rather to the relative relationships among the other components. When interpreting $\tilde{M}_2, \ldots, \tilde{M}_{d-1}$, since they do not contain information about M_1, these variables cannot be directly related to the original M_1. Therefore, if one is interested in the mediation effect of taxa M_ℓ, $\ell \in \{2, \ldots, d\}$, it is necessary to reorder M_ℓ in the ilr transformation so that it becomes the first component as $(M_\ell, M_1, \ldots, M_{\ell-1}, M_{\ell+1}, \ldots, M_d)'$ to interpret the effect of M_ℓ.

For $\ell = 1, \ldots, d$, we propose the following ilr-transformation-based linear mediation models:

$$\begin{aligned} E(Y|X, Z, \tilde{\mathbf{M}}^{(\ell)}) &= \gamma^{(\ell)} X + \beta_1^{(\ell)} \tilde{M}_1^{(\ell)} + \cdots + \beta_{d-1}^{(\ell)} \tilde{M}_{d-1}^{(\ell)} + \mathbf{Z}'\theta^{(\ell)}, \\ E(\tilde{M}_k^{(\ell)}|X, Z) &= \alpha_k^{(\ell)} X + \mathbf{Z}'\eta_k^{(\ell)}, \ k = 1, \ldots, d-1, \end{aligned} \tag{3.11}$$

where $\tilde{\mathbf{M}}^{(\ell)} = (\tilde{M}_1^{(\ell)}, \ldots, \tilde{M}_{d-1}^{(\ell)})'$, and the other notations are defined similarly to those in (3.10).

As previously mentioned, $\alpha_1^{(\ell)} \beta_1^{(\ell)}$ represents an interpretable mediation effect term related to the ℓth taxon, for $\ell = 1, \ldots, d$. To address the multiple testing problem, consider the hypotheses:

$$H_{0\ell} : \alpha_1^{(\ell)} \beta_1^{(\ell)} = 0 \quad \text{versus} \quad H_{A\ell} : \alpha_1^{(\ell)} \beta_1^{(\ell)} \neq 0 \quad \text{for } \ell = 1, \ldots, d.$$

In summary, $\alpha_1^{(\ell)} \beta_1^{(\ell)}$ represents an interpretable mediation effect term for the ℓth taxon. To address the multiple testing problem, the raw P-values for the hypotheses are calculated using the cumulative distribution function of the standard normal distribution. The estimates $\hat{\alpha}_1^{(\ell)}$ and $\hat{\sigma}_{\alpha_1^{(\ell)}}$ are obtained via the OLS method, while $\hat{\beta}_1^{(\ell)}$ and $\hat{\sigma}_{\beta_1^{(\ell)}}$ are derived through the debiased Lasso estimate. This approach allows for the identification of significant mediation effects in high-dimensional settings.

MicroHIMA offers significant advantages in computational efficiency, robustness, and accuracy. It effectively reduces computational time compared to traditional FDR methods and accurately identifies true mediation effects in high-dimensional microbiome data through joint significance tests and the closed testing method. The use of ilr transformation facilitates the application of standard linear models by addressing the compositional nature of the data. However, challenges remain in interpreting transformed mediators, handling data preprocessing requirements, managing complex dependency structures.

3.2.4 PhyloMed

The high dimension and composition nature of the microbiome poses many challenges for mediation analysis. The method proposed in this paper (PhyloMed) takes into account another nature of microbiome, phylogenetic tree, which contains the relationships of Microbial taxa, in addition to the logratio transformation method to handle the composition data. In summary, this phylogeny-based method creates a series of independent local mediation models at the internal nodes of the phylogenetic tree to perform the analysis. Figure 1 shows an example of a simple rooted binary phylogenetic tree to provide a visual illustration. There are 11 microbial taxa on the leaf nodes and 10 internal nodes within the tree.

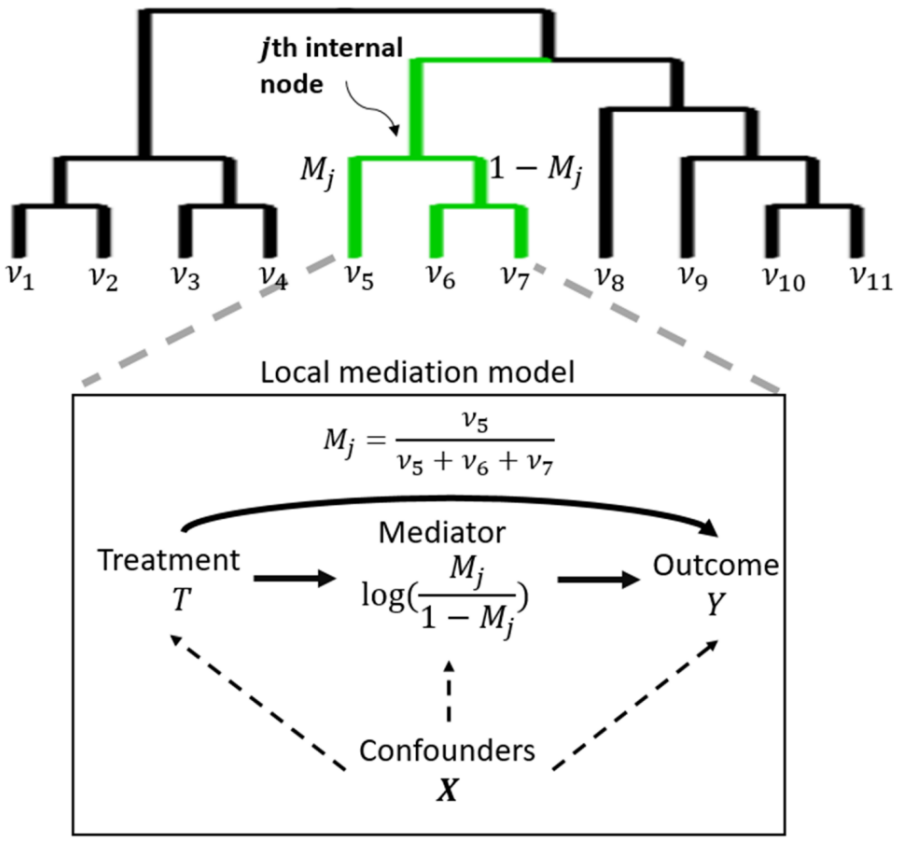

Figure 1: Example of phylogenetic tree and causal path diagram for local mediation model.

Consider the more general case, assuming that there are n subjects, and for each subject $i = 1, \ldots, n$, , let T_i be the treatment variable, Y_i be the outcome variable, $\mathbf{X}_i$ be the set of the confoundeders, and its OTUs correspond to leaf nodes of a rooted phylogenetic tree with J internal nodes. For an internal node j, assume that it has two subnodes, and set the relative abundance of the first and second subnodes to be v_L and v_R, respectively. Then we can get the subcomposition of the internal node j denoted as $(M_{ij}, 1 - M_{ij})$, where $M_{ij} = \frac{v_L}{v_L + v_R}$. Figure 1 shows how to compute subcombination at the internal node j.

This method constructs an independent local mediation model for each subcombination, by using log-ratio transformation. Now log-ratio variable $\log\left(\frac{M_{ij}}{1-M_{ij}}\right)$ acting as mediator, a small pseudocount is added to the data of each leaf node in order to avoid zeros in the transformation. At each internal node j, two regression models are constructed to analyze the mediation effect, denoted as follows:

$$E\left[\log\left(\frac{M_{ij}}{1 - M_{ij}}\right)\right] = \alpha_{jX}^T \mathbf{X}_i + \alpha_j T_i \tag{3.12}$$

$$g\{E(Y_i)\} = \beta_{jX}^T X_i + \beta_{jT}^T T_i + \beta_j \log\left(\frac{M_{ij}}{1 - M_{ij}}\right) \tag{3.13}$$

Equation(3.12) represents the path between the treatment variable T_i and the mediator variable (log-ratio transformed subcomposition), α_j is the the effect of the treatment variable T_i on the mediator

variable. Equation(3.13) represents the path between the mediator variable and the outcome variable Y_i, taking into account the treatment variable T_i, where $g\{\cdot\}$ is a link function chosen based on the type of outcome variable and β_j is the effect of the mediator variable on the outcome variable.

In the local mediation model of jth internal nodes, this method proposes the Composite null hypothesis, defined as follows:

$$H_0^j : \alpha_j \beta_j = 0 \quad \text{versus} \quad H_a^j : \alpha_j \beta_j \neq 0$$

The null hypothesis indicates the absence of mediation effects and can be equated to the joint hypothesis of the three unrelated null hypotheses (i.e., $\alpha_j = 0$ or $\beta_j = 0$ or both are zero). In addition, to improve the accuracy of the test, for each null hypothesis, the method also calculates the p-value for the treatment-mediator association and the mediator-outcome association by using score tests or permutations, respectively, and calculates the final hypothesis by estimating the proportions of the different types of null hypotheses using a mixture of distributions. In addition the test for overall mediation effect was obtained by calculating the harmonic mean of the p-value of each subcomposition mediation model.

The advantage of this method is that it takes into account the fact that the microbiome has the nature of a phylogenetic tree, and by analyzing the subcompositions on the phylogenetic tree, it significantly improves the ability to detect mediating signals, especially when the signals are sparse. At the same time, the disadvantage is that the calculation of this method requires information about the tree and each subcomposition mediation model has to be calculated separately, which adds limitations to the method and the computational effort.

3.2.5 Sparse Microbial Causal Mediation Model (SparseMCMM)

The method is a counterfactual based high-dimensional mediation model for microbiome data. The model to explore the mediation association consisted of two separate models log-contrast regression model and Dirichlet regression model. The authors accommodate the compositional features of microbiome predictors by applying the log-contrast model in the mediator-outcome pathway. In the treatment-mediator pathway, they chose Dirichlet distribution to model the microbial relative abundance. Following the idea of potential outcome in causal inference, the authors defined the total effect, indirect effect and direct effect. They provided tests for both taxon-specific level and community level.

To estimate the direct effect of treatment (T) on the outcome (Y), as well as the mediation effect of the microbiome (M), the authors built the log-contrast model. They picked the $p-th$ taxon as the reference, then the microbial predictors in the model were the ratios of first 1 to $p-1$ taxa compared to the reference taxon.

$$Y_i = \alpha_0 + \alpha_T T_i + \alpha_X^T X_i + \sum_{j=1}^{p-1} \alpha_{Mj} \log\left(\frac{M_{ij}}{M_{ip}}\right) + \sum_{j=1}^{p-1} \alpha_{Cj} T_i \log\left(\frac{M_{ij}}{M_{ip}}\right) + \epsilon_i \tag{3.14}$$

Because the relative abundances are compositional data, they $M_i = (M_{i1}, M_{i2}, \ldots, M_{ip})$ must satisfy the constraint: $\sum_{j=1}^{p} M_{ij} = 1$. α_{Mj}^T are coefficients for each taxon, and α_{Cj}^T are for the interaction terms between taxa and treatment, enforcing that the sum of each coefficient group across all taxa is zero. The estimation were done by least squares method. For regularization, they add lasso penalty to the taxa coefficients and interaction coefficients. L1 norm penalty was also added to the interaction terms.

Dirichlet Regression Model was used to model how treatments and other covariates affect the microbial relative abundance.

$$M_{ij}(T_i, X_i) \sim \text{Dirichlet}(\gamma_1(T_i, X_i), \gamma_2(T_i, X_i), \ldots, \gamma_p(T_i, X_i)) \tag{3.15}$$

and their microbial relative means are linked with treatment and covariates (T_i, X_i) in the generalized linear model fashion with a log link:

$$E[M_{ij}] = \frac{\gamma_j(T_i, X_i)}{\sum_{m=1}^{p} \gamma_m(T_i, X_i)},$$

$$\log\{\gamma_j(T_i, X_i)\} = \beta_{0j} + \beta_{Tj} T_i + \beta_{Xj}^T X_i \tag{3.16}$$

L1 penalty were also added in the maximum likelihood estimation process, so that the specific taxa whose relative abundances are affected by treatment can be selected. Newton-Raphson algorithm was applied to solve the problem.

With the models above, the average causal direct effect of treatment and the average mediation effect of microbiome on the outcome could be determined under the counterfactual framework. Direct Effect (DE):

$$\begin{aligned} DE &= \mathbb{E}[Y_{T=1,M(T=0)} - Y_{T=0,M(T=0)} \mid \mathbf{X}] \\ &= \alpha_T + \alpha_C^T \mathbb{E}[\log(M)|T=0, X] \end{aligned} \tag{3.17}$$

Mediation Effect (ME):

$$\begin{aligned} ME &= \mathbb{E}[Y_{T=1,M(T=1)} - Y_{T=1,M(T=0)} \mid \mathbf{X}] \\ &= \sum_{j=1}^{p} (\alpha_{Mj} + \alpha_{Cj}) \left\{ E\left[\log\left(M_j\right) \mid T=1, \boldsymbol{X}\right] - E\left[\log\left(M_j\right) \mid T=0, \boldsymbol{X}\right] \right\} \end{aligned} \tag{3.18}$$

Total Effect (TE):

$$TE = DE + ME \tag{3.19}$$

They proposed two tests to examine the existence of the microbiome mediation effect on the outcome at the community and taxon levels, denoted as overall mediation effect (OME) and component-wise mediation effect (CME), correspondingly. The null hypothesis of no overall mediation effect at the community level can be expressed $H_0 : ME = 0$. The null hypothesis for the CME test is that none of the individual mediation effects ME_j are significantly different from zero. The null hypothesis is $H_0 : ME_j = 0, \forall j = 1, \cdots, p$, or alternatively $H_0 : \sum_{j=1}^{p} ME_j^2 = 0$. The P-values were obtained using permutation tests.

Pros:

1. The method combines linear log-contrast regression and Dirichlet regression models, capturing complex treatment-mediator-outcome relationships. Dirichlet distribution accomodate the compositional structure of the data.
2. SparseMCMM is able to handle the interactions effect between treatment and microbiome on the outcome.
3. SparseMCMM is designed for high-dimensional data, since they proposed regularization strategy in estimation steps, and the interaction terms when the data suggest that they are absent with the proposed penalized least squares criterion.

Cons:

1. Computational Complexity: SparseMCMM requires extensive computation due to the need for numerous permutations and bootstrap resampling, especially when dealing with large-scale datasets.
2. Dependence on Model Assumptions: Despite being a non-parametric method, SparseMCMM still relies on the assumption of following Dirichlet distribution.
3. Zeros in the data need to be imputed when we fit the log-contrast model.

3.2.6 NPEM, entropy based nonparametric mediation

Nonparametric Entropy Mediation (NPEM, Carter et al. 2020) is a nonparametric framework for mediation analysis of omics data. Like the MODIMA method, NPEM can also capture nonlinear relationships between variables by using information theory and also avoid the assumptions required for regression models. The difference is that this method uses information theory to assess complex relationships between variables rather than simply using distance correlation.

Here the information theory is applied to enable the high-dimensional mediation analysis by utilizing entropy, mutual information and contribution information. Entropy ($H(\cdot)$) measures the uncertainty or information content of a random variable. It is defined as:

$$H(T) = -\sum_{t \in T} p(t) \log p(t)$$

where $p(t)$ represents the probability of observing $T = t$. Mutual Information ($MI(\cdot)$) quantifies the dependence between two random variables. It is defined as:

$$MI(M;Y) = \sum_{m \in M} \sum_{y \in Y} p(m,y) \log \frac{p(m,y)}{p(m)p(y)}$$

or in terms of entropy:

$$MI(M;Y) = H(M) + H(Y) - H(M,Y)$$

Contributed Information ($C(\cdot)$) evaluates the unique contribution of a variable M to Y given a set of other variables T. It is defined as:

$$C(M,Y|T) = MI(M;Y) - \sum_{t \in T} \frac{MI(M;t)}{\|\mathbf{T}\|^2}$$

By utilizing these methods, NPEM is able to capture complex relationship between variables, particularly non-linear and non-additive relationships. Specifically, define the relationship between an exposure variable i and a mediator variable j while controlling for other exposures S as $C(T_i, M_j|S)$, and the relationship between a mediator variable j and the response variable Y while controlling for other mediators K as $C(M_j, Y|K)$. Here, S and K represent subsets of other exposures and other mediators, respectively. Y-i In this article, the authors use Kernel Density Estimation to nonparametrically estimate Mutual Information and Contributed Information metrics. Due to the zero-inflated nature of the microbiome, the concentration of zero counts may lead to problems when using Gaussian kernel density estimation methods, for example, a decrease in variance can lead to narrow bandwidth and overfitting. For this reason, the article proposes two methods which are univariate entropy measure and bivariate entropy measure . The second method can better avoid the potential problems of the kernel density equation. Therefore, we focus on the second method which is based on the first method.

Bivariate entropy measure considers the decomposition of the mediator into presence-absence (whether a taxon unit is detected in the sample or not) and non-zero counts (actual abundance information) so that the contributed information can be estimated respectively. When using kernel density estimates to compute contributed information, it may introduce systematic errors due to the finite sample sizes and bandwidth approximation used to avoid overfitting, defined as expectation bias φ. In order to test whether the relationship between the variables is significant, a general hypothesis can be proposed as:

$$H_0 : MD(\overline{C}) \leq \varphi \quad \text{vs.} \quad H_\alpha : MD(\overline{C}) > \varphi$$

Here, $\overline{C}$ represents the vector that combines the two contributed information, and $MD(\overline{C})$ represents the Mahalanobis distance[**?**] of the contributed information vector. Mahalanobis distance is used to calculate the difference between contributed information vectors. By ensuring that each axis has a mean of zero and a variance of one, it eliminates the effects of correlation and scale and enables more accurate comparisons between different contributed information scores. The Mahalanobis distance is defined as follows: $MD(\overline{C}) = \sqrt{(\overline{C} - \overline{\mu})' \Sigma^{-1} (\overline{C} - \overline{\mu})}$ where $\overline{\mu}$ is the mean vector of $\overline{C}$ and Σ is the covariance matrix of $\overline{C}$. Since the $\overline{C}$ has two dimensions (presence-absence and non-zero counts), the Mahalanobis distance is compared with a chi-square distribution with 2 degrees of freedom to detect abnormally high contributed information values.

Under this hypothesis, accepting the null hypothesis would indicate that the observed contributed information is identical to the expected bias, implying that the relationship between the variables can be attributed to random error or noise. This means that there is no significant relationship between exposure and the mediator or between the mediator and the outcome. On the contrary, if the alternative hypothesis is accepted, it indicates that the relationship between the variables is significant.

The structure of NPEM has more flexibility in addressing non-linear and non-additive relationships, allowing it to handle high-dimensional exposure and mediator variables. Methods are capable of handling continuous, discrete, and mixed data types, which are critical in microbiome and genomics research. However, the NPEM model also has some restrictions. The performance of NPEM depends on the choice of test statistics due to its sensitivity to the proportion of zero values in the data. For example, the choice for univariate and bivariate tests is different according to the characteristics of different data, and it is difficult to choose the appropriate method in practice. It is also worth noting that the Mahalanobis distance indicator does not take direction into account. For example, unusually low signals may be selected.

4 LDM,LDM-Med: Inverse Regression-Based Nonparametric Mediation Analysis Testing Mediation of the Microbiome

Linear Decomposition Model (LDM) is a framework to analyze the associations between microbiome taxa and other covariates based on an inverse regression model. The authors further extended the

framework to microbiome mediation analysis, called LDM-med. Instead of building separate models for treatment-mediator and mediator-outcome pathways, LDM-med explored the mediation effect within a series of taxon-specific linear regression models that include both treatment and outcome as predictors. The responses of the models were the microbiome taxa and the covariates were the residual of treatment, the residual of outcome, and other confounders. Each taxon was considered as the dependent variable in a separate regression model within the overall framework. Therefore, this one-to-one correspondence between taxa and models enables LDM-med to detect mediation effects not only at the community level, but also at the individual taxon level.

4.1 Inverse Regression Model

Start from the following classical models for multiple mediators, for a continuous outcome Y and J continuous mediators with no exposure–mediator or mediator–mediator interactions, the model specifies a linear model for each mediator and a linear model for the outcome that includes the effects of all mediators:

$$E(M_j \mid X, T) = a_{0,j} + a_{1,j}T + a_{X,j}X, \tag{4.1}$$

$$E(Y \mid X, T, M_1, \ldots, M_J) = h_0 + h_X X + h_1 T + \sum_{j=1}^{J} h_{2,j} M_j, \tag{4.2}$$

where X denotes confounding covariates, T denotes the exposure variable, $M_1, \ldots, M_J$ denote the mediators, and Y denotes the outcome variable.

Define T_{rj} to be the residual of T after orthogonalizing against X from (4.1), and Y_r to be the residual of Y after orthogonalizing against (X, T) from (4.2). The inverse regression model for taxon j is

$$E(M_j|Z, T, Y) = b_{0,j} + b_{X,j}X + b_{1,j}T_{rj} + b_{2,j}Y_r. \tag{4.3}$$

The authors showed that $b_{1,j} = a_{1,j}$ and $b_{2,j} = h_{2,j}$. As a result, testing

$$H_{0j} : b_{1,j}b_{2,j} = 0, \tag{4.4}$$

is equivalent to testing $a_{1,j}h_{2,j} = 0$, i.e., whether there exists a mediation effect through taxon j. The test statistics can be obtained by the least-squares estimates from (4.3), denoted by $\hat{b}_{1,j}$ and $\hat{b}_{2,j}$, forming the test statistic as $|\hat{b}_{1,j}\hat{b}_{2,j}|$.

4.2 Testing Mediation Effects

At Individual level, the author provided four different approaches to test the hypothesis. The first approach is following the idea of last section. For taxon j, the test statistic is $U_j = |\hat{b}_{1,j}\hat{b}_{2,j}|$. The distribution under the composite null of no mediation is calculated by breaking the T–M_j association given X and the M_j–Y association given (X, T). The statistic under the b-th permutation is

$$U_j^{(b)} = \max\left(|\hat{b}_{1,j}\hat{b}_{2,j}^{(b)}|, |\hat{b}_{1,j}^{(b)}\hat{b}_{2,j}|, |\hat{b}_{1,j}^{(b)}\hat{b}_{2,j}^{(b)}|\right), \tag{4.5}$$

where $\hat{b}_{1,j}^{(b)}$ and $\hat{b}_{2,j}^{(b)}$ are obtained by permuting T_r and O_r, separately. The permutation P-value for taxon j is then calculated.

The second approach considered P-values $p_{1,j}$ and $p_{2,j}$ for testing $b_{1,j} = 0$ and $b_{2,j} = 0$, respectively. The test statistic $Z_j = \max(p_{1,j}, p_{2,j})$ is taking the larger number of $p_{1,j}$ and $p_{2,j}$, which intuitively means they need both values are small to reject the null. The null distribution of Z_j was also obtained by using the same permutation procedure as in the first approach, so that $Z_j^{(b)} = \min\left(\max(p_{1,j}, p_{2,j}^{(b)}), \max(p_{1,j}^{(b)}, p_{2,j}), \max(p_{1,j}^{(b)}, p_{2,j}^{(b)})\right)$. The third approach is directly implementing the MultiMed (Sampson et al., 2018) procedure that mentioned in other section in this paper to those two p-values $p_{1,j}$ and $p_{2,j}$. In the fourth approach, they employed the HDMT (Dai et al., 2022) procedure to address the P-values $p_{1,j}$ and $p_{2,j}$. See other sections for more details.

At the community level, a global test statistic is constructed by combining those P-values at individual-taxon level. They adopt the Harmonic mean method (Wilson, 2019) to aggregate Z_js, i.e.$J/(\sum_{j=1}^{J} Z_j^{-1})$, where a smaller value corresponds to a stronger evidence against the null hypothesis. Addressing the direction of rejecting the null, a more useful test statistic is the reverse of the Harmonic mean, so the statistic for the global test is $Z_{\text{global}} = \sum_{j=1}^{J} Z_j^{-1}$. The P-value of the test could also be obtained through permutation.

4.3 Discussion

4.3.1 Pros

- LDM-med demonstrates adaptability by accommodating various data types for both exposures and outcomes. This includes continuous variables, discrete or binary variables, multiple exposure or outcome variables, and even survival data (time-to-event outcomes). It is also flexible to include confounders.
- LDM-med provided tests at both individual-taxon level and community level.
- LDM-med method preserves the false discovery rate (FDR) when testing individual taxa and has adequate sensitivity.
- LDM-med could also handle clustered data with the exposure and/or outcome variables varying within the clusters (Zhu et al., 2021), and perform analysis at the presence–absence scale.

4.3.2 Cons

- The inference of LDM-med is based on permutation, which can be computationally expensive in some cases.
- LDM-med doesn't consider the compositional data structure that are widely used in microbiome analysis. Since the linear models are separate from each other, the correlation among the taxa are ignored.

5 IsometricLRTMM, ilr mediation for first taxon

IsometricLRTMM is a mediation analysis frame work based on the structural equation models. The authors built two linear models to capture the relationships among the exposure, the mediators and the outcome. To accommodate the compositional feature of microbiome relative abundance, they considered isometric logratio (ilr) transformation of the relative abundance as the mediator.

5.1 Isometric Logratio Transformation

The relative abundance are compositions that must be non-negative from 0 to 1 and sum up to one. Therefore, it is not suitable to used linear regression to analyze relative abundance directly, since classical linear models are not able to adjust those two constraints. Isometric logratio (ilr) transformation technique proposed by Egozcue et al.(2003) is one of the methods to solve the problem by transforming the compositional data from the simplex S^p to the Euclidean space R^{p-1}. Suppose $\boldsymbol{M} = (M_1, \ldots, M_p)'$ is the vector of relative abundance. The ilr transformation of M_k is

$$\tilde{M}_k = \sqrt{\frac{p-k}{p-k+1}} \ln \frac{M_k}{\left(\prod_{j=k+1}^{p} M_j\right)^{1/(p-k)}} \quad k = 1, \ldots, p-1 \tag{5.1}$$

The transformed first taxon can also be written as

$$\tilde{M}_1 = \frac{1}{\sqrt{p(p-1)}} \left(\ln \frac{M_1}{M_2} + \cdots + \ln \frac{M_1}{M_p} \right) \tag{5.2}$$

which gives us a insight that $\tilde{M}_1$ is formed by a logratio between the compositional part M_1 and the geometric mean of the remaining parts in the composition.

5.2 Structural Equation Model

The authors then put the ilr-transformed coordinates in standard linear regression models in the Euclidean space. Two linear models were built to explore the mediator-outcome and exposure-mediator relationships, respectively.

$$Y = c + \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{p-1} \tilde{M}_{p-1} + \boldsymbol{Z}'\eta + \epsilon \tag{5.3}$$

$$\tilde{M}_k = c_k + \alpha_k X + \boldsymbol{Z}'\theta_k + e_k \quad k = 1, \ldots, p-1 \tag{5.4}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$ are the covariates. α_k represents the relation between X and M_k, and β_k represents the relation between M_k and Y adjusting for the effects of X and $\boldsymbol{Z}$. The interpretation of the parameter β_1 is how much the response variable Y changes averagely by one unit change of the $\tilde{M}_1$, i.e. the logarithm of the ratio between M_1 and the geometric mean of the $M_2, \ldots, M_p$. Given the constraint $\sum_{i=k}^{p} M_k = 1$, it is hard to consider varying one taxon but keeping other taxa unchanged simultaneously. Thus, the authors focused on the relative mediation effect instead of the absolute one for a specific mediator.

5.3 Mediation Effect

Following the idea of SEM framework, the inference of mediation effect can be based on the product of coefficients. But the estimation and inference procedures of IsometricLRTMM was different from the classical SEM after the ilr transformation. The authors showed the test of mediation effect for the first taxon $\tilde{M}_1$ as an example.

$$H_0 : \alpha_1\beta_1 = 0 \quad vs. \quad H_1 : \alpha_1\beta_1 \neq 0$$

De-biased LASSO tenique (Zhang and Zhang, 2014)[**?**] was applied to estimate the coefficients β_k in the first equation to fix the high-dimensional issue. It is proved by Zhang and Zhang that the de-biased LASSO estimator $\hat{\beta}_1$ follows $\left(\hat{\beta}_1 - \beta_1\right)/\sigma_{\beta_1} \xrightarrow{\mathcal{D}} N(0,1)$, so that the p-value be obtained from the asymptotic distribution. σ_{β_1} is the corresponding standard error. More details of estimation were included in the original paper. The joint significance test (JST) was chose to test the existence mediation effect. The p-value of JST was constructed as $P_{joint} = \max\{P_a, P_b\}$, where$P_a = 2\left(1 - \Phi\left(\left|\frac{\hat{\alpha}_1}{\hat{\sigma}_{\alpha_1}}\right|\right)\right)$ and $P_b = 2\left(1 - \Phi\left(\left|\frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}\right|\right)\right)$. We are only able to reject the null when p-values of the coefficients from both models are small enough, which means the coefficients of those two models are non-zero.

5.4 Pros and Cons

5.4.1 Pros

1. The method applied isometric logratio (ilr) transformation to the compositional data, converting the compositions from the simplex space to Euclidean space and making traditional linear regression models applicable.

2. The method employed de-biased Lasso technique to handle the high dimensitonality.

3. The article used the asymptotic distribution of the estimator and proposed a joint significance test method to do the inference.

5.4.2 Cons

1. Based on the characteristics of ilr transformation, only the first ilr transformed variable is interpretable, and other variables are difficult to interpret. Need to reorder the sequence to test other taxa.

2. The method simply replaces zeros with 0.5, which may not be accurate when dealing with a large number of zeros.

6 MicroBVS

The authors proposed a statistical framework to analyze how the microbiome mediated the effect of a treatment on a health outcome via Bayesian method, called MicroBVS. The analysis focused on three components: the binary treatment or exposure (T), the compositional microbial mediators (M) and the continuous health outcome (Y). The authors developed a Bayesian joint model and defined the direct effect and indirect effect based on the model.

6.1 Model

For the Treatment-Mediator pathway the authors apply Dirichlet-multinomial model, and they chose linear regression framework to model the Mediator-Outcome pathway. To link these two models together, they defined balances, a function of microbiome relative abundance, as the predictors of the linear regression model. Balances defined by the authors are the isometric logratio (ilr) transformed relative abundances. Suppose $\boldsymbol{M} = (M_1, \ldots, M_p)'$ is the vector of relative abundance. The ilr transformation of M_j is

$$\tilde{M}_j = \sqrt{\frac{J-j}{p-j+1}} \ln \frac{M_j}{\left(\prod_{k=j+1}^{p} M_k\right)^{1/(J-j)}} \quad j = 1, \ldots, J-1 \tag{6.1}$$

More details about ilr were talked about in the first section of IsometricLRTMM.

The authors assume the microbiome relative abundances $\boldsymbol{M}_i$ for each subject i follow a Dirichlet-multinomial distribution to accommodate the overdispersion of microbiome data and later on facilitate the Markov chain Monte Carlo (MCMC) computation. In this assumption, $\boldsymbol{z}_i \sim \text{Multinomial}\left(\dot{z}_i \mid \boldsymbol{M}_i\right)$, such that $\dot{z}_i = \sum_{j=1}^{J} z_{ij}$, and conjugate priors $\boldsymbol{M}_i \sim \text{Dirichlet}\left(\boldsymbol{\gamma}_i\right)$, where $\boldsymbol{\gamma}_i$ is a J-dimensional vector of concentration parameters. The treatment model is a log-linear regression relating the relative abundances with the treatment.

$$log\left(\gamma_{ij}\right) = \alpha_j + \phi_j T_i + \sum_{p=1}^{P} \theta_{jp} x_{ip}, \tag{6.2}$$

where α_j is a taxon-specific intercept term, ϕ_j is the taxon-specific regression coefficient for treatment, and $\boldsymbol{\theta}_j = \left(\theta_{j1}, \ldots, \theta_{jP}\right)'$ are the taxa-specific regression coefficients. The model cooperates with spike-and-slab priors on each of the regression coefficients ϕ_j and θ_{jp}, with Gaussian slabs of mean 0 and variance r_j^2. They assumed Beta-Bernoulli priors for the latent inclusion indicators, $\varphi_j \sim \text{Beta-Bernoulli}\left(a_v, b_v\right)$ and $\zeta_{jp} \sim \text{Beta-Bernoulli}\left(a_t, b_t\right)$, respectively. In addition, $\alpha_j \sim \text{Normal}\left(0, \sigma_\alpha^2\right)$.

The outcome model is a linear regression model that built the relationship between the outcome and the ilr transformed composition.

$$y_i = c_0 + c_1 t_i + \sum_{j=1}^{J-1} \beta_j \tilde{M}_j + \sum_{p=1}^{P} \kappa_p x_{ip} + \epsilon_i \tag{6.3}$$

The authors made the assumption of the coefficients in the outcome model that have spike-and-slab priors. They constructed the prior construction as $\xi_j \sim \text{Beta-Bernoulli}\left(a_j, b_j\right)$ and $\nu_p \sim \text{Beta-Bernoulli}\left(a_p, b_p\right)$, where (a_j, b_j) and (a_p, b_p) control the sparsity of the balances in the model. c_0, c_1 were assumed to follow $\text{Normal}\left(0, h_c\sigma^2\right)$ and the error term ϵ_i was assumed to follow $\text{Normal}\left(0, \sigma^2\right)$, where $\sigma^2 \sim$Inverse-Gamma(a_0, b_0) for some $a_0 > 0$ and $b_0 > 0$.

6.2 Mediation Effect

Based on the two models above, the direct effect can be defined as

$$\Delta_i = E\left[Y_i\left(T_i = 1, \tilde{\boldsymbol{M}}\left(T_i\right)\right) - Y_i\left(T_i = 0, \tilde{\boldsymbol{M}}\left(T_i\right)\right) \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right] = c_1 \tag{6.4}$$

The direct effect is shared among all subjects, which means Δ_i are the same for all i.

For each subject i, the overall indirect effect of all microbial mediators was defined as

$$\begin{aligned} \delta_i &= E\left[Y_i\left(T_i, \tilde{\boldsymbol{M}}\left(T_i = 1\right)\right) - Y_i\left(T_i, \tilde{\boldsymbol{M}}\left(T_i = 0\right)\right) \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right] \\ &= \sum_{j=1}^{J-1} \beta_j \left(E\left[\tilde{M}_j\left(T_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i\right)\right] - E\left[\tilde{M}_j\left(T_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_i\right)\right]\right) \end{aligned} \tag{6.5}$$

Since the observed covariates are unique for every subject, the indirect effected are subject specific. Given the construction rules of ilr transformation, the taxon-specific indirect effect is a relative one,

which means it depends on the order of the taxa. We showed the definition of mediation effect for the first taxon here, mediation effect of other taxa can also be calculated.

$$\begin{aligned}\delta_{i1} = \beta_1\sqrt{\frac{J-1}{J}}\Big(&E\left[\Psi\left(\gamma_{i1}\left(T_i=1,\boldsymbol{X}_i=\boldsymbol{x}_i\right)\right)-\Psi\left(\sum_{k=1}^{J}\gamma_{ik}\left(T_i=1,\boldsymbol{X}_i=\boldsymbol{x}_i\right)\right)\right]\\&-E\left[\Psi\left(\gamma_{i1}\left(T_i=0,\boldsymbol{X}_i=\boldsymbol{x}_i\right)\right)-\Psi\left(\sum_{k=1}^{J}\gamma_{ik}\left(T_i=0,\boldsymbol{X}_i=\boldsymbol{x}_i\right)\right)\right]\Big),\end{aligned} \tag{6.6}$$

where $\Psi(\cdot)=\frac{d}{dx}\log(\Gamma(x))$ is the digamma function and γ_{ij} is the one they defined in treatment model equation(6.2). The estimation procedure was achieved by MCMC algroithm. The authors stated that only the inference of the first taxon can be directly. Therefore, they suggested to run the algorithm J times and put the different taxa in the first column on each run. Alternatively, the existence of mediation effect can also be tested by constructing the 95% confidence intervals, which can be obtained in MCMC algorithm.

7 MedTest, PCoA mediation analysis for microbiome

7.1 Formula

- Let M be an $n\times m$ count matrix representing the abundances of m OTUs in n samples. - Let T be an $n\times 1$ vector of the independent variable, and Y be an $n\times 1$ vector of the outcome variable. - We assume the microbiome mediates the effect of T on Y through some unknown feature vector $f^{(l)}(M)$.

$$Y = X\gamma' + \epsilon$$
$$f^{(l)}(M) = X\alpha_l + \epsilon_l' \quad (l=1,\ldots,L)$$
$$Y = \sum_{l=1}^{L} f^{(l)}(M)\beta_l + X\gamma + \epsilon''$$

γ',γ represent the total effect and the direct effect of the independent variable T on the outcome Y.
The null hypothesis H_0 can be expressed as

$$H_0 : \alpha_l\beta_l = 0 \text{ for } \forall f_M^{(l)},$$

Apply a distance-based non-parametric method to test the mediation effects. The test consists of two parts: a distance-based test statistic and a permutation scheme to approximate the distribution under the null.

Test Statistic: - Compute residual vectors:

$$X_Z = X - Z(Z^TZ)^{-1}Z^TX$$
$$Y_{X,Z} = Y - [X,Z][X,Z]^TY$$

- Double-center the distance matrix:

$$G = I - \frac{11^T}{n}\cdot A\cdot\left(I-\frac{11^T}{n}\right)$$

where $A=-\frac{1}{2}D^2$, and D is the distance matrix that measures the dissimilarity between the samples based on their microbiota profiles.

- Calculate eigenvectors and eigenvalues:

$$G = U\Lambda U^T$$

Take the first L eigenvectors $(u_1,u_2,\ldots,u_L)$ as microbiome features.

- Calculate the test statistic:

$$T = \sum_{l=1}^{L}\lambda_l|\langle X_Z,u_l\rangle\langle Y_{X,Z},u_l\rangle|$$

Permutation Test: - For B permutations, compute permuted statistics:

$$T_X^{(j)} = \sum_{l=1}^{L} \lambda_l |\langle X_Z^{(j)}, u_l \rangle \langle Y_{X,Z}, u_l \rangle|$$

$$T_Y^{(j)} = \sum_{l=1}^{L} \lambda_l |\langle X_Z, u_l \rangle \langle Y_{X,Z}^{(j)}, u_l \rangle|$$

$$T_{X,Y}^{(j)} = \sum_{l=1}^{L} \lambda_l |\langle X_Z^{(j)}, u_l \rangle \langle Y_{X,Z}^{(j)}, u_l \rangle|$$

- Compute the final permuted statistic:

$$T^{(j)} = \max(T_X^{(j)}, T_Y^{(j)}, T_{X,Y}^{(j)})$$

- Calculate the p-value:

$$p = \frac{1}{B} \sum_{j=1}^{B} I(T^{(j)} \geq T)$$

In this study, multiple distance metrics are used to capture the variation in microbiome composition. These metrics include non-phylogenetic distances (Jaccard distance and Bray-Curtis distance) as well as phylogeny-based UniFrac distances (including unweighted UniFrac, weighted UniFrac, and generalized UniFrac distances).

7.2 Pros and Cons

7.2.1 Pros

1. This method effectively handles high-dimensional, sparse, and non-normal microbiome data, making it suitable for complex microbiome mediation analysis. Method is very general and can be applied to any genomics data with different structures 2. Flexibility: By using multiple distance metrics (including non-phylogenetic and phylogeny-based distances), it captures various types of microbiome variation. 3. Robustness and Power: Simulation studies demonstrate that this method correctly controls Type I error across different mediation models and exhibits strong statistical power in various scenarios.

7.2.2 Cons

Strong Dependence on Model Assumptions: The method assumes linear relationships among variables in the model. Although it can be extended to generalized linear models, it still relies heavily on assumptions such as no unmeasured confounders. Lack of Causal Interpretability: While it can detect mediation effects, this method does not directly provide quantitative estimates of causal relationships, necessitating further experimental validation. Sensitivity to Data Quality: The method requires high-quality data, such as rarefaction to reduce sequencing depth dependency. Poor data quality can affect the reliability of the results.

MedTest assumes that these microbiome features are the units through which the microbiome exert the mediation effect. The power of MedTest may critically depend on whether the exposure– microbiome association and the microbiome–outcome association coincide at the same set of PCs. Furthermore, when the true mediators in the community are rare taxa, the PCs may not effectively capture the variation at these mediators.

7.3 counterfactual assumption

It is a SEM-based model. The linear relationship among variables in the model is assumed.

7.4 Simulation

Data Generation Process for Simulation

In the simulation study of this article, the process of generating the dataset is as follows:

The Dirichlet-Multinomial (DM) model is used to simulate the data, which captures the overdispersion characteristic of real microbiome data. The model parameters are estimated from a real throat microbiome dataset.

Simulation Steps: - Step 1: Generate the independent variable X.

$$X = (x_1, x_2, \ldots, x_n)^T \sim \mathcal{N}(0,1) \text{ (standard normal distribution)}$$

- Step 2: Compute the proportion parameters $p^{(i)}$ for each sample i.

$$p_j^{(i)} = \begin{cases} \frac{e^{(ax_i+\epsilon_i')}p_j}{\sum_{j=1}^{m} e^{(ax_i+\epsilon_i')}p_j} & \text{if } j \in A \\ p_j & \text{if } j \notin A \end{cases}$$

where a is the coefficient representing the relationship between X and the mediating OTUs M, and $\epsilon_i' \sim \mathcal{N}(0,1)$ is the random error. - Step 3: Generate the count matrix M using the above proportion parameters.

$$M \sim \text{Dirichlet-Multinomial}(\text{proportion parameter } p^{(i)}, \text{dispersion } \theta)$$

- Step 4: Compute the outcome variable Y.

$$Y = b \cdot f(P_A) + cX + \epsilon_i''$$

where P_A is the standardized sum of the abundances of the mediating OTUs, and $\epsilon_i'' \sim \mathcal{N}(0,1)$ is the random error.

3. Dataset Dimensions: In the simulated dataset, the sample size $n = 150$, and the number of OTUs m varies across different experiments. Each sample has a sequencing depth of 1000 reads, and the generated count matrix M contains 90 percent zeros, which is similar to the proportion of zeros in real data (approximately 93percent).

In the literature, mediation analysis can be roughly divided into two categories: the structural equation modeling framework [27] and the counterfactual framework [21].

8 IsometricLRTMM, ilr mediation for first taxon

8.1 Simulation

In the simulation study, the data were generated based on the following steps:

1. **Sample size**: The sample size was set to $n = 100$. 2. **Exposure variable**: The exposure variable X was randomly generated from a standard normal distribution $N(0,1)$. 3. **Covariates**: The covariates Z were randomly generated from a standard normal distribution $N(0,1)$. 4. **Mediators**: The mediators $M = (M_1, \ldots, M_p)$ were generated as follows:

- First, the raw values of the mediators M_k were randomly generated from a normal distribution $N(0,1)$.
- Then, the M_k values were transformed using the isometric logratio (ilr) transformation.

5. **Outcome variable**: The outcome variable Y was generated as follows:

- Using the linear regression model $Y = c+\gamma X+M'\beta+Z'\eta+\epsilon$, where the error term ϵ was randomly generated from a normal distribution $N(0,1)$.

Dimensions of Generated Data

In the simulation study, the dimensions of the mediators p were set to different values to test the performance of the method in high-dimensional data. Common values for the dimension include $p = 10, 50, 100$.

9 MODIMA, a Method for Multivariate Omnibus Distance Mediation Analysis

9.1 Formula

Exposure T, mediator M, and response Y 1. Direct effect of exposure on response:

$$Y = i_1 + \gamma X + \epsilon_1 \tag{9.1}$$

2. Effect of exposure on mediator:

$$M = i_3 + \alpha X + \epsilon_3 \tag{9.2}$$

3. Joint effect of mediator and exposure on response:

$$Y = i_2 + \gamma' X + \beta M + \epsilon_2 \tag{9.3}$$

Microbiome analytics must take into account the multivariate nature of such data, and thus often use distance-based approaches. In these cases, power and type I error characteristics are often directly related to the chosen distance metric.

To capture the relationships between multivariate random variables, we use distance correlation (dCor) and partial distance correlation (pdCor):

1. Distance Correlation (dCor):

$$\mathrm{dCor}(X, Y) = \frac{\mathrm{dCov}^2(X, Y)}{\sqrt{\mathrm{dCov}^2(X, X) \cdot \mathrm{dCov}^2(Y, Y)}} \tag{9.4}$$

2. Distance Covariance (dCov):

$$\mathrm{dCov}^2(X, Y) = \frac{1}{n^2} \sum_{i,j} \|X_i - X_j\| \|Y_i - Y_j\| \tag{9.5}$$

3. Partial Distance Correlation (pdCor):

$$\mathrm{pdCor}(X, Y; Z) = \frac{\mathrm{dCor}(X, Y) - \mathrm{dCor}(X, Z) \cdot \mathrm{dCor}(Y, Z)}{\sqrt{1 - \mathrm{dCor}^2(X, Z)}\sqrt{1 - \mathrm{dCor}^2(Y, Z)}} \tag{9.6}$$

The Multivariate Omnibus Distance Mediation Analysis (MODIMA) statistic is defined as:

$$S_d(d_X(X), d_M(M), d_Y(Y)) = \mathrm{dCor}(d_X(X), d_M(M)) \cdot \mathrm{pdCor}(d_X(Y), d_M(M) \mid d_Y(X)) \tag{9.7}$$

Where $d(.)$ denotes the pairwise distance matrices computed from the multivariate observations of exposure X, mediator M, and response Y.

To test the significance of the MODIMA statistic, we use a permutation test. The specific steps are as follows:

1. **Null Hypothesis**: There is no mediation effect between exposure X and response Y, i.e., the observed MODIMA statistic S_d is not significant. 2. **Permutation Steps**: - When the distance correlation between exposure X and mediator M is smaller than the partial distance correlation between mediator M and response Y, permute the rows and columns of the distance matrix $d_X(X)$. - Otherwise, permute the rows and columns of the distance matrix $d_Y(Y)$. 3. **Compute Permutation Statistics**: For each permuted dataset, recompute the MODIMA statistic S_d. 4. **Significance Test**: Compare the observed statistic with the distribution of permuted statistics to compute the p-value:

$$p = \frac{1}{q} \sum_{i=1}^{q} 1(S_d \leq S_d^{(i)}) \tag{9.8}$$

where $S_d^{(i)}$ is the MODIMA statistic for the i-th permutation.

9.2 Pros and Cons

9.2.1 Pros

1. Suitable for High-Dimensional Data: The MODIMA method is particularly suitable for handling high-dimensional data, such as microbiome data, because it uses distance correlation and partial distance correlation, which can effectively capture complex multivariate relationships (from the article, page 2, Introduction section).

2. Power of Energy Statistics: By leveraging partial distance correlation from energy statistics, the method can better capture nonlinear relationships and has higher statistical power in high-dimensional data (from the article, page 3, Motivation for Using Energy Statistics section).

3. Multivariate Analysis: The MODIMA method can handle multivariate exposure-mediator-response triples, providing a more comprehensive analytical framework (from the article, page 4, Multivariate Omnibus Distance Mediation Analysis Statistic section).

4. Open Source Implementation: The method's open-source implementation, including simulation studies and application examples, facilitates use and validation by other researchers (from the article, page 6, Materials and Methods section).

9.2.2 Cons

1. High Complexity: The method is complex, requiring the calculation of distance matrices and partial distance correlations, which can be computationally intensive for large-scale data (from the article, page 5, Empirical Evaluation Simulation section).

2. Time-Consuming Permutation Tests: While accurate, the use of permutation tests to calculate p-values is computationally demanding and time-consuming when a large number of permutations are needed (from the article, page 5, Permutation Testing section).

3. Dependency on Distance Metrics: The results may depend on the chosen distance metric, and different distance metrics may lead to different results, requiring careful selection and interpretation (from the article, page 6, Discussion section).

4. Strict Assumptions: The method's partial distance correlation relies on strict assumptions, particularly when handling non-normal distribution data, which may introduce errors (from the article, page 6, Discussion section). potential pitfall of multivariate mediation analysis pertains to the interpretations of significant mediation results.Further interpretations of this relationship must be treated with caution in order not to attribute this relationship to any individual univariate marginals of the X, M, Y triple, but to treat this relationship as existing in the joint distribution.

Furthermore, the MODIMA paper pointed out a lack of correspondence between conditional independence and zero partial distance correlation, e.g., a non-zero partial correlation in scenarios with conditionally independent variables. It implies that MODIMA may generate false positive findings under the null hypothesis of no mediation, especially when there is a strong direct effect of the exposure on the outcome (θT in model (2)).

9.3 counterfactual assumption

1. Independence Assumption: It is assumed that the relationships between exposure X, mediator M, and response Y can be captured by partial distance correlation. This means that the relationship between exposure and response, given the mediator, can be represented by distance matrices. 2. Suitability of Distance Metrics: It is assumed that the chosen distance metrics (e.g., Euclidean distance, Jensen-Shannon divergence) can effectively capture the correlations in the data and that these distance metrics have consistent statistical properties during computation.

9.4 Simulation

1. Generate normally distributed exposure X and response Y variables. 2. Use saliva and tonsil data from the NIH Human Microbiome Project as mediator M. 3. Generated mediator variables are modeled using a Dirichlet-multinomial distribution. 4. Sample sizes n are set to $\{20, 50, 100, 150\}$.

10 microHIMA, Mediation effect selection in high-dimensional and compositional microbiome data

10.1 Formula

The relative abundances of the high-dimensional microbiome data have an unit-sum restriction, rendering standard statistical methods in the Euclidean space invalid. To address this problem, we use the isometric log-ratio transformations of the relative abundances as the mediator variables. To select significant mediators, we consider a closed testing-based selection procedure with desirable confidence.

Recently, there have been burgeoning statistical or bioinformatical research devoted to studying the mediation effects of microbiome. For instance, Zhang et al22 proposed a distance-based approach for testing the mediation effect of the human microbiome. Sohn and Li23 proposed a sparse compositional mediation model in the simplex space and applied it to a gut microbiome study. Wang et al24 proposed a rigorous sparse microbial causal mediation model for the high-dimensional and compositional microbiome data. Zhang et al25 adopted the isometric log-ratio (ilr)-transformation and debiased Lasso techniques to develop a joint significance test for the mediation effect of human gut microbiome with a focus on prespecified taxa. In this work, we propose a novel method to select mediating microbial taxa. As existing methods for microbiome mediation analysis are primarily designed to test the overall mediation effect (eg, Sohn and Li23 and Wang et al24), our major contribution is to propose a statistical procedure **to select individual taxa that mediate the path between exposure and phenotype.**

Log-ratio transformation for the relative abundances part same as IsometricLRTMM

High-dimensional inference part same as IsometricLRTMM

In this section, we focus on developing a selection procedure for mediation effects. Denote by $\{H_{0\ell}\}_{\ell=1}^{d}$ the collection of hypotheses of interest (elementary hypotheses) in Equation (4), $T_0 \subseteq \{1,\ldots,d\}$ is the index set of true null hypotheses. The closed testing methods consider not only the elementary hypotheses, but also all intersection hypotheses of the form $H_A = \bigcap_{i\in A} H_{0i}$, where $A \subseteq \{1,\ldots,d\}$ and $A \neq \emptyset$. An intersection hypothesis H_A is true if and only if H_i is true for all $i \in A$. Let $\mathcal{A}$ be the collection of all subsets of $\{1,\ldots,d\}$. We define $\mathcal{U}_{0.05}$ as the collection of all $A \in \mathcal{A}$ such that H_A is rejected, where $P(T_0 \notin \mathcal{U}_{0.05}) \geq 0.95$ (or equivalently $P(T_0 \in \mathcal{U}_{0.05}) \leq 0.05$). For any $S \subseteq \{1,\ldots,d\}$ of selected hypotheses to reject (referred to as discoveries), Goeman et al provided a simultaneous 0.95-confidence lower-bound, $LB_{0.05}(S)$, for the number of true discoveries in S. Below we summarize the procedure of Goeman et al in Algorithm 1.

Algorithm 1. Closed Testing-Based Algorithm

Step 1: Obtain the raw (unadjusted) P-values $P_1,\ldots,P_d$ for the elementary hypotheses $H_{01},\ldots,H_{0d}$. Sort these P-values in the increasing order as $P_{r1} \leq P_{r2} \leq \cdots \leq P_{rd}$, where $r_i \in \{1,\ldots,d\}$. Let $R_i = \{r_1,\ldots,r_i\}$ be the set of the smallest i P-values. Similarly define $R_{d-i} = \{r_1,\ldots,r_{d-i}\}$ to be the set of the smallest $d-i$ P-values. Denote by $L_i = \{1,\ldots,d\} \setminus R_{d-i}$ the set of the largest i P-values.

Step 2: Calculate the 0.95-confidence lower-bound for the number of true discoveries in S,

$$LB_{0.05}(S) = \max_{1\leq k\leq |S|} 1 - k + |\{i \in S : h_{0.05}P_i \leq 0.05k\}|,$$

where $|S|$ denotes the number of elements in S, and $h_{0.05} = \max\{1 \leq i \leq d : L_i \notin \mathcal{U}_{0.05}\}$ denotes the size of the set of largest P-values not rejected.

Algorithm 2. Mediation Effect Selection Algorithm

Step 1: Sort the P-values $\{P_{\text{raw},\ell}\}_{\ell=1}^{d}$ in (6) as $P_{\text{raw},r1} \leq P_{\text{raw},r2} \leq \cdots P_{\text{raw},rd}$. Let $S_1 = \{r_1\}, S_2 = \{r_1, r_2\}, \cdots, S_K = \{r_1, r_2, \cdots, r_K\}$, and $K = \max\{i : P_{\text{raw},ri} \leq 0.05, i = 1, \cdots, d\}$.

Step 2: Run Algorithm 1 to obtain the values of $LB_{0.05}(S_k)$ in (7), for $k = 1, \cdots, K$.

Step 3: Define $J_1 = LB_{0.05}(S_1)$ and $J_k = LB_{0.05}(S_k) - LB_{0.05}(S_{k-1})$, where $k = 2, \cdots, K$. The estimated index set of significant mediators is

$$\hat{\mathcal{S}} = \{r_k : J_k = 1, for k = 1, \cdots, K\}.$$

Remark 1: In Algorithm 2, sorting the P-values $\{P_{\text{raw},\ell}\}_{\ell=1}^{d}$ can save computation time as only K values of $LB_{0.05}(S)$ are calculated. Otherwise, a total of d lower-bounds $LB_{0.05}(S)$ need to be calculated. Hence, our method has a computational advantage, especially when the value of K/d is small.

In summary, we first impose the ilr-transformation on the relative abundances, then we refit the ilr-transformed variables in the linear mediation models. Because only the first element of the ilr-transformed variables is interpretable, we permute the orders of the original d mediators in turn to ensure that each taxon should play the role of the first element. In the structural equation modeling

(SEM) framework, we obtain the raw P-values by joint significant tests for the component-wise mediation effects. Furthermore, we apply a novel closed testing-based selection method for the ilr-transformed high-dimensional mediators.

10.2 Pros and Cons

10.2.1 Pros

1. **Effectiveness and Applicability**: The proposed method shows good effectiveness and applicability through simulations and real data applications, as demonstrated by the application to murine gut microbiome data, verifying the practical utility of the method [Chapter 5]. 2. **Computational Efficiency**: Compared to traditional FDR methods, this method saves computational time by avoiding the need to adjust all hypothesis tests individually. This is especially evident when the number of hypothesis tests is large [Section 2.3, Algorithm 2]. 3. **Robustness**: The method effectively handles the high-dimensional and compositional nature of microbiome data. The use of ilr transformation removes the compositional structure of the data, allowing standard linear models to be applied directly [Section 2.1]. 4. **Accuracy**: Through joint significance tests and the closed testing method, this approach excels in selecting significant mediators, accurately identifying true mediation effects [Chapter 3].

10.2.2 Cons

1. **Interpretation Difficulty**: Since only the first element of the ilr-transformed variables is interpretable, the interpretation of other transformed mediators in the model is not straightforward [Section 2.1]. 2. **Dependence on ilr Transformation**: Although ilr transformation solves the analysis problem of compositional data, it may require additional preprocessing of the data, such as handling zeros or outliers, in some cases [Chapter 5, Conclusion]. 3. **Complex Dependency Structures**: While the method can handle high-dimensional data, it may still face limitations when dealing with complex dependency structures, such as interactions within the microbiome [Chapter 5, Conclusion]. 4. **High Computational Burden**: Despite its computational efficiency advantage, the method may still have a computational burden when dealing with very high-dimensional data, especially when multiple cross-validations are needed to determine penalty parameters Section 2.2].

10.3 counterfactual assumption

SEM-based

1. **Independently and Identically Distributed (i.i.d.) Assumption**: It is assumed that the observations (Y_i, X_i, Z_i, M_i) are independently and identically distributed (i.i.d.), meaning that each sample is independent of others and follows the same distribution Section 2.2. 2. **Linear Relationship Assumption**: It is assumed that there is a linear relationship between the exposure X, covariates Z, mediators $\tilde{M}$, and the response Y in the high-dimensional linear mediation model. The model is as follows:

$$E(Y|X, Z, \tilde{M}) = \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{d-1}\tilde{M}_{d-1} + Z'\theta,$$
$$E(\tilde{M}_k|X, Z) = \alpha_k X + Z'\eta_k, \quad k = 1, \ldots, d-1$$

[Section 2.1]. 3. **Centering Assumption**: It is assumed that the response Y and the ilr-transformed mediators $\tilde{M}_k$ are centered, meaning their means are zero Section 2.1. 4. **Normality Assumption**: When performing debiased Lasso estimation, it is assumed that the residuals are normally distributed. The formula is as follows:

$$\hat{\sigma}_{\beta_1^{[\ell]}} = \frac{1}{\sqrt{n}}\hat{\sigma}_{\epsilon^{[\ell]}} \sqrt{\frac{\sum_{i=1}^n (R_i^{[\ell]})^2}{n}} \left| \frac{\sum_{i=1}^n R_i^{[\ell]} \tilde{M}_{i1}^{[\ell]}}{n} \right|$$

where $(\hat{\sigma}_{\epsilon^{[\ell]}})^2$ is the variance of the residuals Section 2.2.

10.4 Simulation

In the simulation study, the data generation process is as follows:

1. **Sample Size and Dimensions**:

- The sample size n is not explicitly mentioned, but it can be assumed to be a reasonable number such as 100 or 200 based on context.
- The dimension d, i.e., the number of mediators, is set accordingly.

2. **Compositional Operators and Power Transformation**:

- Two compositional operators are introduced, defined as:

$$\eta \oplus \zeta = \left(\frac{\eta_1 \zeta_1}{\sum_{j=1}^{d} \eta_j \zeta_j}, \ldots, \frac{\eta_d \zeta_d}{\sum_{j=1}^{d} \eta_j \zeta_j} \right)'$$

$$\eta' = \left(\frac{\eta_1^r}{\sum_{j=1}^{d} \eta_j^r}, \ldots, \frac{\eta_d^r}{\sum_{j=1}^{d} \eta_j^r} \right)'$$

3. **Data Generation Model**:

- Data is generated from the following compositional mediation model:

$$\mathbf{M} = \mathbf{m}_0 \oplus a^X \oplus \mathbf{e}$$

$$Y = c_0 + cX + (\log \mathbf{M}')'\mathbf{b} + \epsilon$$

where $\mathbf{M}$ is the mediator matrix, $\mathbf{m}_0$ is the baseline composition, $\mathbf{e}$ is noise generated from a multivariate logistic normal distribution with mean zero and covariance matrix Σ_e, and ϵ is noise following a normal distribution .

4. **Exposure and Parameters**:

- The exposure variable X is generated from a standard normal distribution $N(0, 1)$, and a^X is generated according to formula (10) .
- Parameters c_0 and c are set to 1 and 0.5, respectively.

5. **Special Settings**:

- **Case I**: The parameter $\mathbf{b}$ is set to (1.3, -0.7, -0.6, 0, ..., 0)', where the first 3 elements correspond to significant mediators. $\mathbf{a}_i = \frac{13\mathbf{u}_i}{6\mathbf{u}_i'\mathbf{u}_i}$ and $\mathbf{u}_i$ is generated from the uniform distribution $U(0, 1)$. The covariance matrix $\Sigma_e = 2\mathbf{N}$, where $\mathbf{N}$ is the identity matrix plus the product of a column vector of 1's and a row vector of 1's .
- **Case II**: Similar to Case I, but $\mathbf{e}$ is generated from a t-distribution with 3 degrees of freedom .

11 LDM,LDM-Med, Inverse Regression-Based Nonparametric Mediation Analysis Testing Mediation of the Microbiome

11.1 Simulation

In the simulation study, data are generated based on upper-respiratory-tract microbiome data (Charlson et al., 2010). The generation process is as follows:

1. **Setting Baseline Relative Abundance**: First, set the baseline relative abundances $p_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})$ for all taxa, estimated from the population means.
2. **Inducing the Effects of Exposure on Taxa**: For unexposed samples, keep p_i unchanged; for exposed samples, decrease p_{ij} for some associated taxa by a percentage, and redistribute the decreased amount evenly over the remaining associated taxa.
3. **Introducing Sample Heterogeneity**: Introduce sample heterogeneity by drawing the sample-specific composition p_i from the Dirichlet distribution with mean p_i.
4. **Generating Count Data**: Generate count data using the Multinomial distribution with the mean p_i and the sequencing depth sampled from a normal distribution.

5. **Generating Outcome Variable**: The outcome variable is generated by the mediator and type-II null taxa, using the model:

$$O_i = \beta_{TO} T_i + \beta_{MO} \text{scale}\left(\sum_{j \in M_1} p_{ij} - \sum_{j \in M_2} p_{ij}\right) + \alpha_{MO} \text{scale}\left(\sum_{j \in N_1} p_{ij} - \sum_{j \in N_2} p_{ij}\right) + \epsilon_i,$$

where ϵ_i is drawn from a normal distribution.

The simulated data include 100 samples and 856 taxa. The sample size of 30 is also considered to match the murine microbiome dataset.

12 PERMANOVA-med, Extension of PERMANOVA to Testing the Mediation Effect of the Microbiome

12.1 Formula

PERMANOVA [7] is currently the most commonly used distance-based method in analysis of microbiome data. Although it was originally developed for testing microbiome associations, we find that we can extend PERMANOVA to testing microbiome mediation effects by using the idea of inverse regression and including both the exposure and the outcome as covariates whose F statistics capture the exposure–microbiome association and the microbiome–outcome association conditional on the exposure, respectively.

Motivation toward Inverse Regression

Inverse regression is a commonly used approach to testing associations. It has the key advantage of accommodating different types of outcome variables, including multivariate variables. Assuming the relationships among the exposure (T), mediator (M), outcome (O), and confounders (Z) are depicted in Figure 1b. In the classical mediation model, we typically specify the following forward regression models:

$$E(M|Z,T) = \alpha_0 + \alpha_Z Z + \alpha_T T \tag{12.1}$$

$$E(O|Z,T,M) = \theta_0 + \theta_Z Z + \theta_T T + \theta_M M \tag{12.2}$$

However, for complex microbiome data, such forward outcome models are not easily generalizable. To overcome this limitation, we adopt the inverse regression model:

$$E(M|Z,T,O) = \beta_0 + \beta_Z Z + \beta_T T_r + \beta_O O_r \tag{12.3}$$

where T_r and O_r denote the residuals of T and O after orthogonalization against Z. Model (3) implies that by testing $\beta_T \beta_O = 0$, we can test for the existence of a mediation effect.

Overview of PERMANOVA

PERMANOVA is based on a linear model of covariates that partition a given distance matrix along each covariate. Specifically, when the Euclidean distance measure is used on the relative abundance data, PERMANOVA partitions the total variance of the relative abundance data across all taxa into variance explained by each covariate. Using our implementation in permanovaFL, the design matrix X is grouped into K submodels, i.e., $X = (X_1, X_2, \ldots, X_K)$. Each submodel includes components that will be tested jointly, such as a single covariate, multiple covariates, or multiple indicators for a categorical covariate. The submodels are first processed into sequentially orthogonal, unit vectors by the Gram-Schmidt process, so that the partition of the distance matrix is unambiguous.

Let Δ denote the $n \times n$ distance matrix, which is often Gower-centered. PERMANOVA tests the effect of the k-th submodel by using the F statistic:

$$F_k \propto \frac{\text{Tr}[X_k X_k^T \Delta_k X_k X_k^T]}{\text{Tr}[(I - \sum_{k' \neq k} X_{k'} X_{k'}^T) \Delta_k (I - \sum_{k' \neq k} X_{k'} X_{k'}^T)]} \tag{12.4}$$

PERMANOVA assesses the significance of the F statistic via permutation.

PERMANOVA-med: Extension of PERMANOVA to Mediation Analysis

Under model (3), we set submodels $X_1 = Z$, $X_2 = T_r$, and $X_3 = O_r$, and denote the PERMANOVA F statistics for testing microbiome associations with T_r and O_r by F_T and F_O, respectively. We propose to test the existence of a mediation effect by the microbiome using the test statistic:

$$U_{\text{PERMANOVA-med}} = F_T \times F_O \tag{12.5}$$

To claim a mediation effect by the microbiome, both the exposure-microbiome and microbiome-outcome associations (given the exposure) are required to be significant. Accordingly, we construct the following statistic for the b-th permutation replicate:

$$U^{(b)}_{\text{PERMANOVA-med}} = \max\{F_T^{(b)} F_O, F_T F_O^{(b)}, F_T^{(b)} F_O^{(b)}\} \tag{12.6}$$

The p-value is obtained as the proportion of $U^{(b)}_{\text{PERMANOVA-med}}$ that are equal to or larger than the observed statistic $U_{\text{PERMANOVA-med}}$.

12.2 Pros and Cons

12.2.1 Pros

1. **Wide Applicability**: PERMANOVA-med is applicable to various types of microbiome mediation analysis, including multivariate exposures, continuous, binary, and multivariate outcome variables, as well as survival outcomes. These features are mentioned in the introduction (page 2, paragraph 1).

2. **Flexibility in Adjusting Confounders**: PERMANOVA-med allows for the adjustment of confounders, which is crucial in complex microbiome data analysis. This is emphasized in the methods section (page 4, paragraph 1).

3. **Omnibus Test**: PERMANOVA-med can provide an omnibus test that combines multiple distance matrices, which is particularly important when selecting the optimal distance measure (page 4, paragraph 2).

4. **Control of Type I Error**: Simulation studies indicate that PERMANOVA-med performs well in controlling Type I error, as verified in the results section (page 9, paragraph 2).

5. **Handling of Rare Taxa**: PERMANOVA-med excels in handling data that include rare taxa (page 12, paragraph 2).

12.2.2 Cons

1. **Limited to Community-Level Mediation Effect Detection**: PERMANOVA-med can only detect mediation effects at the community level and cannot identify specific mediating taxa (page 16, paragraph 1).

2. **Need for External Information for Causal Interpretation**: Although PERMANOVA-med can detect associations, declaring true mediation effects requires additional information on causal direction (page 16, paragraph 2).

3. **Computational Complexity**: PERMANOVA-med involves multiple submodels and permutation tests, leading to high computational complexity, especially with large-scale data (page 3, paragraph 1).

4. **Dependence on Data Structure**: The effectiveness of PERMANOVA-med relies heavily on the structure of the distance matrices of the data. Complex data structures might impact the method's effectiveness (page 4, paragraph 2).

12.3 counterfactual assumption

SEM-based

1. **Linear Model Assumption**: It is assumed that the relationships among the exposure (T), mediator (M), outcome (O), and confounders (Z) can be described by linear models. This is mentioned when introducing the inverse regression model (page 4, paragraph 1).

2. **Orthogonalization Assumption**: It is assumed that the confounders (Z), exposure (T), and outcome (O) can be orthogonalized to obtain the orthogonalized residuals T_r and O_r. This is detailed in the description of the inverse regression model (page 4, paragraph 1).

3. **Correctness of Distance Matrices**: It is assumed that the distance matrices Δ accurately reflect the dissimilarities among samples and that these distance matrices can be Gower-centered. This is mentioned in the description of the PERMANOVA method (page 4, paragraph 2).

4. **Independently and Identically Distributed (i.i.d.) Assumption**: It is assumed that the samples are independently and identically distributed, which is a common assumption in permutation tests. This is implicitly mentioned in the description of the permutation tests (page 5, paragraph 1).

12.4 Simulation

The method for generating data and the dimensions of the data in the simulation study are as follows:

1. **Data Generation Method**: - **Based on Real Data**: The simulation study is based on upper respiratory tract (URT) microbiome data, which includes 856 taxa (page 8, paragraph 1). - **Exposure Variables**: - **Binary Exposure**: Samples are randomly divided, with half of the samples $T_i = 1$ and the other half $T_i = 0$ (page 8, paragraph 2). - **Continuous Exposure**: Samples T_i are drawn from a Beta(2, 2) distribution (page 8, paragraph 2). - **Mediator Variables**: Sample-specific composition vectors π_i are generated using a Dirichlet distribution, and read count data are generated using a Multinomial distribution. Baseline relative abundances are adjusted to reflect the effect of the exposure on the mediator taxa (page 8, paragraph 3). - **Outcome Variables**: - For M-common and M-mixed, the mediator taxa influence the outcome through their relative abundances or presence-absence statuses. The outcome variable O_i is generated using a linear model (page 9, paragraph 1). - **Confounders**: In settings with a binary exposure, a binary confounder Z_i is generated (page 9, paragraph 2).

2. **Data Dimensions**: - **Number of Samples** (n): 100 or 200 (page 8, paragraph 2). - **Number of Taxa**: 856 taxa (page 8, paragraph 1).

13 PhyloMed: a phylogeny-based test of mediation effect in microbiome

13.1 Formula

Te distance-based tests can achieve good power when many microbial taxa in the community mediate the treatment efect on the outcome. However, their power is limited when the number of mediating taxa is small. LDM-med [19] tests the mediation efect at individual taxa using the relative abundance and combine test statistics across taxa to produce a global test. False discovery rate (FDR)-controlling procedures are applied to identify mediating taxa.

In this article, we develop a phylogeny-based mediation analysis method (PhyloMed) for the high-dimensional mediator of microbial composition. PhyloMed models the microbiome mediation efect through a cascade of independent local mediation models of subcompositions on the internal nodes of the phylogenetic tree.

We develop a testing procedure to ensure the PhyloMed local mediation test p-values are asymptotically mutually independent and uniformly distributed under the global null hypothesis that no mediation efect in any of the internal nodes (Methods). We apply a FDR-controlling procedure to the local mediation p-values and identify mediating internal nodes.

PhyloMed Framework

We consider a random sample of n subjects measured on a set of operational taxonomic units (OTUs). Suppose the OTUs are placed on the leaves of a rooted phylogenetic tree with J internal nodes. For subject $i = 1, \ldots, n$, let $(M_{ij}, 1 - M_{ij})$ be the subcomposition at the jth internal node. In the PhyloMed framework, instead of having a single model on the composition of all leaf-level OTUs, we build a collection of independent local models, each for a subcomposition on a particular internal node of the tree. This modeling approach is similar to a Polya tree process and has been adopted by several methods for microbiome differential abundance testing. Despite the independence of subcompositions M_{ij}'s over internal nodes, this modeling approach allows a rich dependence structure among leaf-level OTUs.

We apply the log-ratio transformation to the subcomposition and use the log-ratio variable (i.e., $\log\left(\frac{M_{ij}}{1-M_{ij}}\right)$) as the mediator. We need to handle zeros in the log-ratio transformation. In each local mediation model at an internal node, we remove subjects with both components of the subcomposition being zero as they carry no information on the subcomposition. We add 0.5 to the counts aggregated to the left and right child nodes for the remaining subjects. Although adding a small value (pseudocount) is a simple and commonly used practice to avoid zeros in the log transformation, the choice of pseudocount is arbitrary, and there is no consensus on the optimal value. Studies have shown that the pseudocount approach can lead to biased normalization, and downstream data analysis can be sensitive to the choice of pseudocount.

For each subject i, let T_i be the treatment variable, Y_i be the outcome variable, and X_i be a set of confounders that may affect the treatment, mediator, and outcome. The causal path diagram of the local mediation model at the jth internal node is represented as follows:

1. Mediation Model:

$$E\left(\log \frac{M_{ij}}{1-M_{ij}}\right) = \alpha_{jT}T_i + \alpha_{jX}X_i$$

2. Outcome Model:

$$g\{E(Y_i)\} = \beta_{jX}X_i + \beta_{jT}T_i + \beta_j \log\left(\frac{M_{ij}}{1-M_{ij}}\right)$$

where $g(\cdot)$ is the link function depending on the type of outcome. Since M_{ij}'s over internal nodes are independent, we can fit J low-dimensional models $[Y_i \mid X_i, T_i, \text{logit}(M_{ij})]$ as Eq. (2) instead of a large joint model $[Y_i \mid X_i, T_i, \text{logit}(M_{i1}), \ldots, \text{logit}(M_{iJ})]$ for the purpose of hypothesis testing.

The potential outcomes framework has established a series of identifiability assumptions such that models (1) and (2) lead to quantification of the causal mediation effect. The rigorous definition of causal mediation using the potential outcomes framework is provided in Supplementary Information Note A.1.

Composite Null Hypothesis Tests in Local Mediation Models

In each local mediation model, we are interested in testing whether the subcomposition at the jth internal node lies in the causal pathway from the treatment to the outcome. For both continuous and binary outcomes, the null and alternative hypotheses for this testing problem can be formulated as follows:

$$H_0^j : \alpha_j\beta_j = 0 \quad \text{versus} \quad H_a^j : \alpha_j\beta_j \neq 0$$

The null hypothesis can be equivalently expressed as the union of three disjoint null hypotheses:

$$H_{00}^j : \alpha_j = 0, \beta_j = 0$$
$$H_{10}^j : \alpha_j \neq 0, \beta_j = 0$$
$$H_{01}^j : \alpha_j = 0, \beta_j \neq 0$$

Sobel's test and the joint significance test have been widely applied to test the mediation null hypothesis. However, these tests have severely deflated type I error and lack power because they fail to take into account the composite nature of the null hypothesis.

To address this issue, several new mediation tests were recently developed. In PhyloMed, we adopt a mixture-distribution-based approach to handle the composite null hypothesis and produce well-controlled type I error for multiple local mediation tests. Let $P_{\alpha j}$ and $P_{\beta j}$ denote the p-values for testing $\alpha_j = 0$ and $\beta_j = 0$, respectively. We define the mediation test statistic for H_0^j as:

$$P_{\max j} = \max(P_{\alpha j}, P_{\beta j})$$

The mediation test p-value at the jth node can be estimated as follows:

$$\Pr(P_{\max j} \leq p_{\max j}) = \pi_{00}p_{\max j}^2 + \pi_{10}p_{\max j}\Pr(P_{\alpha j} \leq p_{\max j} | \alpha_j \neq 0) + \pi_{01}p_{\max j}\Pr(P_{\beta j} \leq p_{\max j} | \beta_j \neq 0)$$

We employ nonparametric estimates to estimate $\Pr(P_{\alpha j} \leq p_{\max j} | \alpha_j \neq 0)$ and $\Pr(P_{\beta j} \leq p_{\max j} | \beta_j \neq 0)$. Finally, the mediation test p-value at the jth node can be estimated as:

$$p_j = \pi_{00}p_{\max j}^2 + \pi_{10}p_{\max j}\Pr(P_{\alpha j} \leq p_{\max j} | \alpha_j \neq 0) + \pi_{01}p_{\max j}\Pr(P_{\beta j} \leq p_{\max j} | \beta j \neq 0)$$

Mediating Subcomposition Detection

We deduce that under the global null hypothesis H_0, the mediation test statistics $P_{\max j}$'s of internal nodes are asymptotically mutually independent as the sample size goes to infinity.

Under H_0^j, the two power function probabilities $\Pr(P_{\alpha j} \leq t_j | \alpha j \neq 0)$ and $\Pr(P_{\beta j} \leq t_j | \beta j \neq 0)$ converge to 1. Therefore, based on formula (3), $\Pr(P_{\max j} \leq t_j)$ can be approximated by:

$$\Pr(P_{\max j} \leq t_j) \approx \pi_{00}t_j^2 + \pi_{10}t_j + \pi_{01}t_j$$

Therefore, we can apply the standard Benjamini-Hochberg (BH) procedure to identify nodes with significant mediation effects on the phylogenetic tree.

Global Mediation Test

We can combine all subcomposition mediation test p-values to test the global mediation null hypothesis that there is no mediation effect in any of the internal nodes (i.e., $H_0 : \bigcap_{j=1}^{J} H_0^j$). Here, we employ the harmonic mean p-value (HMP) method. The HMP is defined as:

$$\hat{p} = \frac{J}{\sum_{j=1}^{J} \frac{1}{p_j}}$$

The global test p-value can be obtained by calculating the tail probability from the $\hat{p}$'s null distribution.

13.2 Pros and Cons

13.2.1 Pros

1. **Increased Discovery Power**: - PhyloMed, by analyzing subcompositions on the phylogenetic tree, significantly improves the discovery power for mediation signals, especially when the signals are sparse (Page 4, Paragraph 2):

2. **Good Control of Type I Error**: - By using a mixture distribution to handle the composite null hypothesis, PhyloMed ensures that the local mediation test p-values are asymptotically independent and uniformly distributed under the global null hypothesis (Page 14, Paragraph 2):

3. **Handling High-dimensional Compositional Data**: - PhyloMed effectively deals with the compositional and high-dimensional nature of microbiome data by applying log-ratio transformation and adding pseudocounts (Page 14, Paragraph 1):

4. **Multiple Testing Control**: - PhyloMed employs the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR) in multiple testing, ensuring accurate identification of significant mediation nodes (Page 16, Paragraph 2):

5. **Independent Local Model Construction**: - PhyloMed constructs independent local models at each internal node of the phylogenetic tree, using the relative abundance ratio of subcompositions as mediators. This method avoids the difficulty of jointly modeling a large number of microbial taxa and fully leverages the evolutionary relationships of microbes (Page 13, Paragraph 2):

Tis tree-guided approach efectively enriches mediation signals that tend to cluster on the phylogenetic tree and boosts the power of the test for weak mediation efects among taxa. Moreover, PhyloMed accounts for the compositional nature of the relative abundance data and the composite mediation null hypothesis, resulting in well-calibrated p-values for testing local subcomposition mediation efects.

13.2.2 Cons

1. **Dependence on Phylogenetic Tree**: - The performance of PhyloMed may be affected if the phylogenetic tree is misspecified, potentially reducing the effectiveness of mediation signal detection (Page 12, Paragraph 2):

2. **Computational Complexity**: - While PhyloMed has advantages in handling high-dimensional data, its complex computational steps may increase the computational burden, especially with large-scale data (Page 17, Paragraph 1):

Te assumption of no unmeasured confounding is critical in obtaining an unbiased estimate of the mediation efect and establishing the causal interpretation in mediation analysis.

Power can be afected if the phylogenetic tree is misspecifed. For instance, if the mediation taxa are clustered on the true tree but more scattered on the misspecifed tree, signals on the internal nodes may become less condensed and more challenging to detect.

13.3 counterfactual assumption

1. **Assumption of No Unmeasured Confounding**: - It is assumed that there are no unmeasured confounding factors present, meaning all possible confounders affecting the treatment, outcome, and mediator variables are measured and controlled in the model (Page 11, Paragraph 1)

2. **Assumption of Linear Relationships**: - It is assumed that there are linear relationships among the variables in the mediation model and the outcome model. For the mediation model, the linear relationship is reflected in the log-ratio transformation of the subcomposition; for the outcome model, the linear relationship is between the outcome variable and the mediator and treatment variables (Page 13, Paragraph 2)

3. **Assumption of Independence**: - It is assumed that the local mediation models at different internal nodes of the phylogenetic tree are independent. This means each local mediation model can be estimated separately without being influenced by other nodes (Page 13, Paragraph 2)

13.4 Simulation

To resemble reality, the simulated data is based on a real microbiome dataset containing samples from 900 healthy subjects. In each simulation, 50 or 200 subjects are randomly sampled and divided into treatment and control groups

The top 100 most abundant OTUs and their associated phylogenetic tree were used. The phylogenetic tree contains 99 internal nodes.

Under the null hypothesis of no mediation effect, the sets of treatment-associated OTUs S_α and outcome-associated OTUs S_β do not overlap. Under the alternative hypothesis, $S_\alpha = S_\beta$ and both sets of OTUs are set as mediators.

Generation of Treatment and Outcome Variables For each treatment-associated OTU $k \in S_\alpha$, it is randomly decided whether to change the abundance of the OTU in the treatment group or the control group. For each subject in the chosen group, the abundance is increased and the outcome variable is generated. The continuous outcome variable is generated using a linear log-contrast regression model, and the binary outcome variable is generated using a logistic log-contrast regression model.

14 Sparse Microbial Causal Mediation Model (SparseMCMM)

The method is a counterfactual based high-dimensional mediation model for microbiome data. The model to explore the mediation association consisted of two separate models log-contrast regression model and Dirichlet regression model. The authors accommodate the compositional features of microbiome predictors by applying the log-contrast model in the mediator-outcome pathway. In the treatment-mediator pathway, they chose Dirichlet distribution to model the microbial relative abundance. Following the idea of potential outcome in causal inference, the authors defined the total effect, indirect effect and direct effect. They provided tests for both taxon-specific level and community level.

14.1 Model

To estimate the direct effect of treatment (T) on the outcome (Y), as well as the mediation effect of the microbiome (M), the authors built the log-contrast model. They picked the $p-th$ taxon as the reference, then the microbial predictors in the model were the ratios of first 1 to $p-1$ taxa compared to the reference taxon.

$$Y_i = \alpha_0 + \alpha_T T_i + \alpha_X^T X_i + \sum_{j=1}^{p-1} \alpha_{Mj} \log\left(\frac{M_{ij}}{M_{ip}}\right) + \sum_{j=1}^{p-1} \alpha_{Cj} T_i \log\left(\frac{M_{ij}}{M_{ip}}\right) + \epsilon_i \tag{14.1}$$

Because the relative abundances are compositional data, they $M_i = (M_{i1}, M_{i2}, \ldots, M_{ip})$ must satisfy the constraint: $\sum_{j=1}^{p} M_{ij} = 1$. α_{Mj}^T are coefficients for each taxon, and α_{Cj}^T are for the interaction terms between taxa and treatment, enforcing that the sum of each coefficient group across all taxa is zero.

Dirichlet Regression Model was used to model how treatments and other covariates affect the microbial relative abundance.

$$M_{ij}(T_i, X_i) \sim \text{Dirichlet}(\gamma_1(T_i, X_i), \gamma_2(T_i, X_i), \ldots, \gamma_p(T_i, X_i)) \tag{14.2}$$

and their microbial relative means are linked with treatment and covariates (T_i, X_i) in the generalized linear model fashion with a log link:

$$\begin{aligned} E[M_{ij}] &= \frac{\gamma_j(T_i, X_i)}{\sum_{m=1}^{p} \gamma_m(T_i, X_i)}, \\ \log\{\gamma_j(T_i, X_i)\} &= \beta_{0j} + \beta_{Tj} T_i + \beta_{Xj}^T X_i \end{aligned} \tag{14.3}$$

With the models above, the average causal direct effect of treatment and the average mediation effect of microbiome on the outcome could be determined under the counterfactual framework. Direct Effect (DE):

$$\begin{aligned} DE &= \mathbb{E}[Y_{T=1, M(T=0)} - Y_{T=0, M(T=0)} \mid \mathbf{X}] \\ &= \alpha_T + \alpha_C^T \mathbb{E}[\log(M) | T = 0, X] \end{aligned} \tag{14.4}$$

Mediation Effect (ME):

$$ME = \mathbb{E}[Y_{T=1,M(T=1)} - Y_{T=1,M(T=0)} \mid \mathbf{X}] \tag{14.5}$$

$$ME = (\alpha_M^T + \alpha_C^T)(\mathbb{E}[\log(M)|T=1,X] - \mathbb{E}[\log(M)|T=0,X]) \tag{14.6}$$

$$ME = \sum_{j=1}^{p} ME_j \tag{14.7}$$

Total Effect (TE):

$$TE = DE + ME \tag{14.8}$$

We propose two tests to examine whether the microbiome has any mediation effect on the outcome or not, at the community and taxon levels, denoted as OME and CME. Since the null hypothesis of no overall mediation effect at the community level can be expressed $H_0 : ME = 0$.

- Overall Mediation Effect (OME):

$$\text{OME} = (\alpha_M^T + \alpha_C^T)(\mathbb{E}[\log(M)|T=1,X] - \mathbb{E}[\log(M)|T=0,X]) \tag{14.9}$$

The null hypothesis for the CME test is that none of the individual mediation effects ME_j are significantly different from zero. Mathematically, this is expressed as:

$$H_0 : ME_j = 0 \quad \text{for all } j \in \{1,\ldots,p\} \tag{14.10}$$

An equivalent formulation of the null hypothesis is:

$$H_0 : \sum_{j=1}^{p} ME_j^2 = 0 \tag{14.11}$$

The P-values are estimated using permutation tests.

14.2 Pros and Cons

14.2.1 Pros

1. SparseMCMM is designed specifically for high-dimensional and compositional microbiome data, capable of effectively managing a large number of variables. This is particularly useful for modern microbiome studies that often deal with high-dimensional datasets. By using sparse regression techniques, SparseMCMM can select important variables from high-dimensional data, reducing noise and overfitting.

2. Flexible Model Structure: The method combines linear log-contrast regression and Dirichlet regression models, capturing complex treatment-mediator-outcome relationships. with Dirichlet regression, SparseMCMM can handle the interaction between treatment and microbiome with relation to the outcome in a more flexible manner through the proposed regularization strategy, which addresses concerns regarding the potential bias caused by neglecting the presence of interaction effects

3. Moreover, SparseMCMM can automatically drop the interaction terms when the data suggest that they are absent with the proposed penalized least squares criterion; CMM lacks such flexibility. Thirdly, SparseMCMM selects casual taxa with regularization techniques, while CMM identifies the key taxa based on confidence interval estimates.

14.2.2 Cons

1. Computational Complexity: SparseMCMM requires extensive computation due to the need for numerous permutations and bootstrap resampling, especially when dealing with large-scale datasets.

2. Dependence on Model Assumptions: Despite being a non-parametric method, SparseMCMM still relies on specific model assumptions, such as log-contrast regression and Dirichlet distribution. If these assumptions do not hold, the model estimates may be affected. Therefore, it is crucial to carefully validate these assumptions when applying SparseMCMM.

3. One limitation of SparseMCMM is that it can only take a single time point of microbiome data into the proposed framework.

14.3 counterfactual assumption

Four sufficient identifiable assumptions. There are four assumptions must be satisfied in order to identify the causal effect of treatment on outcome using the proposed causal mediation models under the counterfactual framework (VanderWeele and Vansteelandt, 2009, 2014; VanderWeele, 2016; Huang and Pan, 2016). First, no unmeasured confounders for the relationship between treatment and outcome, i.e.,

$$Y(t, \mathbf{m}) \perp\!\!\!\perp T \mid \mathbf{X}$$

for all levels of t and $\mathbf{m}$; secondly, no unmeasured confounders for the relationship between mediator and outcome, i.e.

$$Y(t, \mathbf{m}) \perp\!\!\!\perp \mathbf{M} \mid T, \mathbf{X}$$

for all levels of t and $\mathbf{m}$; thirdly, no unmeasured confounders for the relationship between treatment and mediator, i.e.

$$\mathbf{M}(t) \perp\!\!\!\perp T \mid \mathbf{X}$$

for all levels of t; lastly, there is no unmeasured confounders for the relationship between mediator and outcome that can be affected by the treatment, i.e.

$$Y(t, \mathbf{m}) \perp\!\!\!\perp \mathbf{M}(t^*) \mid \mathbf{X}$$

for all levels of t, t^* and $\mathbf{m}$.

14.4 Simulation

Generate the Treatment T: For n total subjects, randomly assign 50 percent to the treatment group ($T = 1$) and the others to the control group ($T = 0$). The sample sizes used for estimation are $n = 50$, 100, 300, and 500; for testing, $n = 100$.

Generate the Microbiome M: - Simulate microbiome data based on the Dirichlet distribution, reflecting the real microbial composition. The mean relative abundances of p taxa for the treatment and control groups follow:

$$E[M_{ij}] = \frac{\gamma_j(T_i)}{\sum_{m=1}^{p} \gamma_m(T_i)}, \quad \log \gamma_j(T_i) = \beta_{0j} + \beta_{Tj} T_i$$

Here, $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ represents the log-transformed baseline relative abundances for p taxa, set as the corresponding estimates from the actual data using the R package 'dirmult'. The values of p used in the simulations are 10, 25, and 50.

Generate the Outcome Y: Generate the outcome Y based on the treatment T and microbiome composition M. The model is fitted as:

$$Y_i = \alpha_0 + \alpha_T T_i + \sum_{j=1}^{p-1} \alpha_{Mj} \log\left(\frac{M_{ij}}{M_{ip}}\right) + \sum_{j=1}^{p-1} \alpha_{Cj} T_i \log\left(\frac{M_{ij}}{M_{ip}}\right) + \epsilon_i$$

The fitting parameters are set as follows: intercept $\alpha_0 = 0$, treatment effect $\alpha_T = 1$, non-causal taxa effects $\alpha_{Mj} = \alpha_{Cj} = 0$. For pr causal taxa, the regression coefficients for the main effect and interaction effect are α_M and α_C.

15 NPEM, entropy based nonparametric mediation

15.1 Formula

Information theory compares joint distributions of two or more variables with the marginal distributions of subsets to measure association between variables. This can capture nonlinear and non-additive associations by observing changes in distribution of the outcome as compared to distance based and regression modeling approaches which can only capture linear association with the outcome.

The information can be measured using Shannon Entropy and Mutual Information . Shannon entropy represents the uncertainty, potential information, from a discrete random variable or random vector, and is defined as the amount of information produced by a stochastic process:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{15.1}$$

where $p(x)$ represents the probability of observing $X = x$. Shannon entropy of a multivariate process between two variables X and Y can be calculated using joint Shannon entropy:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \tag{15.2}$$

where $p(x, y)$ represents the probability of observing $X = x$ and $Y = y$.

Mutual information (MI) is defined as the overlap of information produced by multiple stochastic processes:

$$MI(X; Y) = H(Y) + H(X) - H(X, Y) \tag{15.3}$$

$$MI(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{15.4}$$

To capture the unique mutual information from a variable X, we additionally define the contributed information to be the mutual information of one variable given a set of measured variables W:

$$C(X, Y, W) = MI(X; Y) - \sum_{w \in W} MI(X; w) \tag{15.5}$$

In the mediation model, α represents the effect coefficient of the exposure variable T on the mediator variable M. β represents the effect coefficient of the mediator variable M on the response variable Y, when the exposure variable T is controlled. γ represents the direct effect coefficient of the exposure variable T on the response variable Y, without passing through the mediator variable M. γ' represents the total effect of the exposure variable T on the response variable Y, including both the direct effect and the indirect effect through the mediator variable M. β_1 represents the overlapping part of the effect coefficient of the mediator variable M on the response variable Y, which is also contained in the exposure variable T. β_2 represents the unique effect coefficient of the mediator variable M on the response variable Y, which is not contained in the exposure variable T.

When $\beta_2 \neq 0$, it indicates the presence of a mediation effect. If $\beta_2 = 0$, there are two possible scenarios: 1. If $\beta_1 = 0$ and $\beta_2 = 0$, then M does not provide any information about Y, indicating no mediation effect. 2. If $\beta_1 \neq 0$ and $\beta_2 = 0$, all information provided by M is also contained in T, indicating perfect collinearity, and thus no conclusion can be drawn about the existence of mediation effects.

We propose two approaches for mediation testing using mutual information.

Univariate Entropy Measure

For a particular taxon j to be a mediating taxon, there must be significant relationships from at least one gene through it to the response. For each taxon j the hypotheses are:

$$H_0 : C(T_i, M_j, S) \leq \phi_{\alpha,j}, \forall i \in \{1, \ldots, I\} \quad \text{OR} \quad C(M_j, Y, T) \leq \phi_{\beta 2}$$

$$H_a : \exists i \in \{1, \ldots, I\} : C(T_i, M_j, S) > \phi_{\alpha,j} \quad \text{AND} \quad C(M_j, Y, T) > \phi_{\beta 2}$$

The parameters $\phi_{\alpha,j}$ and $phi_{\beta 2}$ represent the expected bias for contributed information with a fixed taxon j and Y respectively.

Bivariate Entropy Measure

To account for the difference in scale and correlation between presence-absence and nonzero counts, we will utilize Mahalanobis distance (Mahalanbois, 1936):

$$MD(C^*) = \sqrt{(C^* - \mu^*)^T S^{-1} (C^* - \mu^*)}$$

where μ^* represents the vector of means for C^* and S represents the covariance of the two contributed information scores in C^*.

15.2 Pros and Cons

15.2.1 Pros

This method is a nonparametric method which does not require linear and additive assumptions about the data, making it more flexible in handling nonlinear and non-additive relationships.

It can handle high-dimensional exposure and mediator variables, which is difficult for traditional regression methods.

Capable of handling continuous, discrete, and mixed data types, which is crucial in bioinformatics and genomics research.

Simulation studies show that this method has higher power and lower error rates in detecting significant mediation effects compared to existing methods.

15.2.2 Cons

Sensitivity to Zero-Inflated Data. When dealing with count data with a large number of zeros, special handling may be required to avoid bias.

Dependence on Kernel Density Estimation. In some cases, kernel density estimation may lead to overfitting, especially when the data sample size is small.

15.3 counterfactual assumption

This is a nonparametric approach that avoids many of the assumptions required for traditional regression models.

15.4 Simulation

1. Gene Expression Data: Data for 300 genes is generated using a normal distribution. The first 150 genes are generated with a standard deviation of 0.5, and the second 150 genes with a standard deviation of 2.0.
2. Taxon Count Data: Taxon counts are generated using a negative binomial distribution with excess zeros added. The probability of excess zeros is weighted by the log ratio of abundance to population mean.

1. **Simulation Study One**: - Sample Size: 40 and 80 per group - Excess Zero Probability: High (80- Signal Strength: 50- Total of four data scenarios

a total of 20 datasets are generated and evaluated

signal strength is defined as the ratio of the average difference between the healthy and diseased groups to the standard deviation of the noise. This definition is used to measure the strength of the true signal in the data. The specific formula is:

$$\text{signal strength} = \frac{d}{s}$$

where d represents the average difference between the healthy and diseased groups, and s represents the standard deviation of the noise

16 Hypothesis testing for mediation effect in high-dimensional genomics studies

16.1 Formula

In this paper, we propose a new method to test the mediation effect with high-dimensional continuous mediators. The methods section is divided into the following parts:

Causal Mediation Model

We propose two regression models to represent the causal mediation model. In the first model, the outcome Y_i is determined by covariates X_i (with the first element being 1 for the intercept), exposure S_i, mediators G_i, and possible interactions S_iG_i:

$$Y_i = X_i^T\beta_X + S_i\beta_S + G_i^T\beta_G + S_iG_i^T\beta_C + \epsilon_{Y_i},$$

where $\epsilon_{Y_i} \sim N(0, \sigma^2)$, $\beta_G = (\beta_{G1}, \ldots, \beta_{Gp})^T$, and $\beta_C = (\beta_{C1}, \ldots, \beta_{Cp})^T$.

The j-th mediator is determined by covariates X_i and exposure S_i:

$$G_{ji} = X_i^T \alpha_{Xj} + S_i \alpha_{Sj} + \epsilon_{Gji},$$

or equivalently, one can specify a multivariate model for the mediators:

$$G_i = AX_i + S_i \alpha_S + \epsilon_{G_i},$$

where $\epsilon_{G_i} \sim N_p(0, \Sigma)$, and A and α_S are parameter matrices.
The mediation effect (ME) can be defined with counterfactual notation as follows:

$$\text{ME} = E[Y(s_1, G(s_1))|X] - E[Y(s_1, G(s_0))|X].$$

Transformation of Mediators

We use spectral decomposition to transform the mediators G into uncorrelated variables P, simplifying the model computation. For the covariance matrix Σ, there exists an orthogonal matrix U such that $U\Sigma U^T$ is diagonal. The transformed model is:

$$P_i = A^* X_i + S_i \alpha_S^* + \epsilon_{P_i},$$

where $P_i = UG_i$, $A^* = UA$, and $\alpha_S^* = U\alpha_S$.
The transformed model can be rewritten as:

$$Y_i = X_i^T \beta_X + S_i \beta_S + P_i^T \beta_G^* + S_i P_i^T \beta_C^*.$$

Hypothesis Testing Procedure

To test the significance of the mediation effects, we propose three types of hypothesis tests: 1. **Marginal Mediation Effect** ψ:

$$H_0 : \psi = \alpha_S^T (\beta_G + \beta_C) = 0.$$

2. **Component-Wise Mediation Effect** δ:

$$H_0 : \delta_j = \alpha_{Sj}(\beta_{Gj} + \beta_{Cj}) = 0, \quad \forall j.$$

3. **L2 Norm of Component-Wise Effects** τ:

$$H_0 : \tau = \|\delta\|_2 = 0.$$

To approximate the distribution of these tests, we use a Monte-Carlo procedure to generate independent parameter estimates and calculate the corresponding statistics. By repeatedly sampling, we obtain the empirical distribution of the test statistics, which allows us to compute p-values and conduct hypothesis testing.

16.2 Pros and Cons

16.2.1 Pros

1. **Handling High-Dimensional Data**: The method is capable of handling situations with a large number of mediators by reducing computational complexity through spectral decomposition and transformation models. This is detailed in Section 2 "The Causal Mediation Model" and Section 3.

2. **Improved Estimation Accuracy**: By transforming high-dimensional mediators into uncorrelated variables, the method allows for more accurate estimation of mediation effects in low-dimensional regression models. This is specifically explained in Section 3.

3. **Applicability to Small Sample Sizes**: The method performs well when dealing with datasets with high-dimensional mediators and small sample sizes, using transformation models to avoid numerical instability issues encountered in traditional methods. This is mentioned in Sections 3 and 4.

4. **Multiple Hypothesis Tests**: The method proposes three types of hypothesis tests: marginal mediation effect, component-wise mediation effects, and the L2 norm of component-wise effects, offering multiple ways to evaluate the significance of mediation effects. This is discussed in detail in Section 4.

16.2.2 Cons

1. **Dependence on Normality Assumption**: The method assumes that the mediators follow a normal distribution, which may not always hold in practice. Violation of this assumption could affect the accuracy of the estimates and tests. This is mentioned in Section 7.

2. **Interpretability Limitations**: The transformed mediators P_j are linear combinations of the original mediators, so the component-wise mediation effects δ_j may lack intuitive biological interpretation. This is detailed in Section 7 "Discussion".

3. **Potential Assumption Violations**: Although the method accounts for multiple mediators, it assumes no unmeasured confounders. If these assumptions are violated, the results might be biased. This is discussed in Section 2.1 "Identifiability and Model Assumptions".

16.3 counterfactual assumption

The method proposed in this paper requires several key counterfactual assumptions when establishing the causal mediation model and conducting hypothesis testing for mediation effects. These assumptions ensure the identifiability and accuracy of the model estimates. The main counterfactual assumptions include:

1. **No Confounding for the Exposure-Outcome Relationship**: Assuming no confounding between the exposure S and the outcome Y after adjusting for covariates X, i.e., $Y(s) \perp S|X$.

2. **No Confounding for the Mediator-Outcome Relationship**: Assuming no confounding between the mediators G and the outcome Y after adjusting for exposure S and covariates X, i.e., $Y(s,g) \perp G|S,X$.

3. **No Confounding for the Exposure-Mediator Relationship**: Assuming no confounding between the exposure S and the mediators G after adjusting for covariates X, i.e., $G(s) \perp S|X$.

4. **No Mediator-Outcome Confounders Affected by the Exposure**: Assuming there are no confounders for the mediator-outcome relationship that are themselves affected by the exposure S, i.e., $Y(s,g) \perp G(s')|X$.

16.4 Simulation

In the simulation studies of this paper, the methods for generating data and the dimensions of the data are as follows:

1. **Simulation for Continuous Outcome Variables**: - Sample size $n = 500$ and $n = 50$. - Number of mediators $p = 50$. - Exposure variable S takes values 0, 1, and 2, randomly assigned to approximately equal numbers of subjects. - Mediators G are generated according to the model $G_{ji} = X_i^T \alpha_{Xj} + S_i \alpha_{Sj} + \epsilon_{Gji}$, where α_{Sj} ranges from 0 to 0.15 or from 0 to 0.4 in different simulation settings, and ϵ_G follows a multivariate normal distribution with mean 0 and covariance matrix Σ. - Outcome variable Y is generated according to the model $Y_i = X_i^T \beta_X + S_i \beta_S + G_i^T \beta_G + S_i G_i^T \beta_C + \epsilon_{Y_i}$, where ϵ_{Y_i} follows a standard normal distribution $N(0,1)$.

2. **Simulation for Dichotomous Outcome Variables**: - Sample size $n = 2000$ and $n = 400$. - Number of mediators $p = 50$. - Exposure variable S takes values 0, 1, and 2, randomly assigned to approximately equal numbers of subjects. - Mediators G are generated according to the same model as for continuous outcome variables. - Outcome variable Y is generated according to the model $\text{logit}[P(Y_i = 1|S_i, G_i, X_i)] = X_i^T \beta_X + S_i \beta_S + G_i^T \beta_G + S_i G_i^T \beta_C$, with the intercept set to -0.35.

17 CCMM

17.1 Formula

proposed a causal composition mediation model (CCMM) specifically for microbiome mediators which utilized a bootstrap covariance matrix to perform log-contrast compositional regression. While these approaches may avoid concerns associated with th

In this paper, we contribute to extending the applicability of mediation analysis further by proposing an estimating method for the causal direct and indirect effects when mediators are compositional.

Our framework utilizes two components: (1) an estimation method based on the compositional operators of Aitchison (1982) and the composition algebra of Billheimer, Guttorp and Fagan (2001) and (2) the linear log contrast regression of Lin et al. (2014), Shi, Zhang and Li (2016).

Compositional Mediation Model and Causal Interpretation

For unit i, let T_i be the treatment, M_i be a vector of k compositional mediators, Y_i an outcome, and X_i a set of pre-treatment variables that may affect the treatment, mediator, and outcome. We adopt the potential outcomes framework for model assumptions and identification of causal direct and indirect effects. Let $M_i(t)$ be the potential outcome under $T_i = t$ and $Y_i(t, m)$ the potential outcome under $T_i = t$ and $M_i = m$. Thus, we can express an observed variable as $M_i = M_i(T_i)$ and similarly $Y_i = Y_i(T_i, M_i(T_i))$.

Suppose we have a random sample of size n from a population where we observe Y_i, T_i, X_i, and M_i for each unit i. Note that $M_i \in S_{k-1}$ for all i, i.e., $M_i = \{(M_{i1}, \ldots, M_{ik}) : M_{ij} > 0, j = 1, \ldots, k, \sum_{j=1}^{k} M_{ij} = 1\}$. We propose the following compositional mediation model:

$$M_i = (m_0 \oplus aT_i \oplus h_{X_{i1}} \oplus \cdots \oplus h_{X_{iq}}) \oplus U_{1i}$$

$$Y_i = c_0 + cT_i + (\log M_i)^\top b + X_i^\top g + U_{2i} \quad \text{subject to } b^\top 1_k = 0$$

where m_0 is the baseline composition, c_0 is the baseline for Y_i, a, b, c are path coefficients, $h_1, \ldots, h_q$ and g are nuisance parameters corresponding to X_i, and 1_k is a vector of k ones. Note that $U_{1i} \in S_{k-1}$ as $M_i \in S_{k-1}$. We assume U_{1i} is perturbed around the identity element J_{k-1} of S_{k-1}, i.e., $E(U_{1i}) = Jk-1$, where $J_{k-1} = 1_k/k$. We also assume $U_{2i} \sim N(0, \sigma^2)$.

Model Assumptions and Identification

Identification of the causal direct and indirect effects requires the following assumptions:

$$\{Y_i(t', \log m), \log M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

$$Y_i(t', \log m) \perp\!\!\!\perp \log M_i(t) \mid T_i = t, X_i = x$$

No interaction between T_i and M_i on the response.

These assumptions are an extension of the sequential ignorability assumptions for the single mediator model.

Parameter Estimation, Variance Estimation, and Tests of Mediation Effects

Estimation of Composition Parameters and Covariance Matrix

To estimate the parameters in the model, we propose the following objective function:

$$(a, h_r, m_0) = \arg \min_{a, h_r, m_0 \in S_{k-1}} \sum_{i=1}^{n} \|M_i \ominus (m_0 \oplus aT_i \oplus h_{X_{i1}} \oplus \cdots \oplus h_{X_{iq}})\|^2$$

The objective function is not convex in terms of a_j, m_{0j}, h_{rj} but is in terms of $\text{alt}(a)_j, \text{alt}(m_0)_j, \text{alt}(h_r)_j$. The optimal solution can be obtained by solving the following system of linear equations with constraints $m_0, a, h_r \in S_{k-1}$:

where $\zeta_0 j = k \sum_{i=1}^{n} \log M_{ij} - \sum_{k'=1}^{k} \sum_{i=1}^{n} \log M_{ik'}, \zeta_1 j = k \sum_{i=1}^{n} T_i \log M_{ij} - \sum_{k'=1}^{k} \sum_{i=1}^{n} T_i \log M_{ik'}, \xi_r j = k \sum_{i=1}^{n} X_{ir} \log M_{ij} - \sum_{k'=1}^{k} \sum_{i=1}^{n} X_{ir} \log M_{ik'}$, and for any $\nu, D(\nu)$ is defined as:

$$D(\nu) = \begin{bmatrix} (k-1)\sum_{i=1}^{n} \nu_i & -\sum_{i=1}^{n} \nu_i & \cdots & -\sum_{i=1}^{n} \nu_i \\ -\sum_{i=1}^{n} \nu_i & (k-1)\sum_{i=1}^{n} \nu_i & \cdots & -\sum_{i=1}^{n} \nu_i \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{i=1}^{n} \nu_i & -\sum_{i=1}^{n} \nu_i & \cdots & (k-1)\sum_{i=1}^{n} \nu_i \end{bmatrix}$$

The covariance matrix of the estimated parameters $\hat{\Sigma_a}$ is obtained using the percentile method of Machado and Parente (2005) from the bootstrap distribution of $\hat{a}$.

Estimation of Compositional Regression Parameters and Covariance Matrix

We use the linear log-contrast model and the debias procedure to estimate regression parameters (i.e., b and c) and their covariance matrix. Specifically, we first solve the following objective function:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \left(\sum_{i=1}^{n} \left(Y_i - cT_i - (\log M_i)^\top b - X_i^\top g \right)^2 \right) + \lambda \|\beta\|_1$$

where $\beta = (c, b, g)^\top$, and λ is a tuning parameter. We then apply the debias procedure to the solution of the objective function to obtain unbiased estimates and their covariance matrix.

Hypothesis Test of Mediation Effect

The null hypothesis of no total compositional mediation effect is given by:

$$H_0 : (\log a)^\top b = 0$$

The null hypothesis of no component-wise mediation effect is given by:

$$H_0 : \log(ka_j)b_j = 0 \quad \forall j \in \{1, 2, \ldots, k\}$$

We propose two approaches to test these null hypotheses: an extension of the Sobel test and a bootstrap approach. In testing the null hypothesis with the former, the square root of the first order asymptotic variance of the total indirect effect is computed with the estimated covariance matrices of $\log(k\hat{a})$ and $\hat{b}$ and used as a standard error of the total indirect effect in the Z-test. To avoid the assumption of normality for the indirect effect, we use a bootstrap approach.

17.2 Pros and Cons

17.2.1 Pros

1. **Handling Sparsity**: The introduction of the sparse compositional mediation model helps address the common issue of sparsity in high-dimensional data. This makes it easier to extract meaningful information from microbiome data that contains many irrelevant or redundant variables (Chapter 2, "Compositional Mediation Model and Causal Interpretation"). 2. **Causal Effect Estimation**: By incorporating compositional algebra into causal mediation analysis, the model can more accurately estimate causal direct and indirect effects. This is particularly important for understanding the complex relationships between the microbiome and health (Chapter 2, "Compositional Mediation Model and Causal Interpretation"). 3. **Multiple Hypothesis Testing Methods**: The method provides multiple ways to test for causal mediation effects, including an extension of the Sobel test and a bootstrap approach, increasing the flexibility and applicability of the method (Chapter 3, "Parameter Estimation, Variance Estimation, and Tests of Mediation Effects").

17.2.2 Cons

1. **Model Complexity**: The sparse compositional mediation model is complex, involving many parameters and assumptions, which can lead to high computational demands and time consumption in practical applications (Chapter 2, "Compositional Mediation Model and Causal Interpretation"). 2. **Strict Assumptions**: The model relies on strict assumptions, such as no interaction assumptions, which may limit its applicability to certain real-world datasets. If these assumptions are not met, the model's estimates may be unreliable (Chapter 2, "Model Assumptions and Identification"). 3. **Need for Sensitivity Analysis**: Although sensitivity analysis methods are provided, they can be complex and require substantial computational resources, posing higher demands on data processing and result interpretation (Chapter 3, "Parameter Estimation, Variance Estimation, and Tests of Mediation Effects").

17.3 counterfactual assumption

1. **Sequential Ignorability Assumption**: It is assumed that given covariates X_i, the treatment T_i is independent of the potential mediator M_i and the outcome Y_i. This means that T_i is not influenced by unobserved factors, making the assignment of T_i random conditional on X_i.

Mathematically:

$$\{Y_i(t', \log m), \log M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

2. **No Interaction Assumption**: It is assumed that there is no interaction effect between the mediator M_i and the treatment T_i on the outcome Y_i. This means that Y_i is simply a linear combination of T_i, M_i, and covariates X_i without complex interaction effects.

Mathematically:

$$Y_i(t', \log m) \perp\!\!\!\perp \log M_i(t) \mid T_i = t, X_i = x$$

3. **No Unmeasured Confounding Assumption**: It is assumed that all potential confounders are observed and measured, and included in the covariates X_i. This ensures that, conditional on X_i, there is no unmeasured confounding between the treatment T_i, the mediator M_i, and the outcome Y_i.

17.4 Simulation

1. **Generation of Covariates**: Generate X_i as covariates, assuming X_i follows a standard normal distribution, i.e., $X_i \sim N(0, 1)$.

2. **Generation of Treatment**: Generate the treatment variable T_i, assuming T_i is a binary variable with distribution $T_i \sim Bernoulli(0.5)$.

3. **Generation of Mediators**: - First, generate the latent mediator U_{1i}. - Generate the mediator M_i using the sparse compositional model $M_i = (m_0 \oplus aT_i \oplus h_{X_{i1}}) \oplus U_{1i}$, where m_0 is the baseline composition, and a and h are the corresponding path coefficients.

4. **Generation of Outcome**: Generate the outcome Y_i according to the following model:

$$Y_i = c_0 + cT_i + (\log M_i)^\top b + X_i^\top g + U_{2i}$$

where $U_{2i} \sim N(0, \sigma^2)$ and $b^\top 1_k = 0$.

In the simulation study, the authors consider different data dimensions to evaluate the performance of the method:

1. **Sample Size** n: Different sample sizes are used to assess the model's performance under small and large samples. Specific sample sizes are provided with multiple settings in the paper.

2. **Number of Mediators** k: The number of mediators in the simulated data k also varies, with typical mediator numbers including $k = 3, 5, 10$, etc.

3. **Number of Covariates**: The dimension of X_i is denoted as p, where p represents the number of covariates.

18 MedZIM: Mediation analysis for Zero-Inflated Mediators with applications to microbiome data

18.1 Formula

To address the issue of mediation analysis in the context of zero-inflated data structures, we developed a novel mediation analysis method called MedZIM. This method is based on the potential outcomes framework and is designed for zero-inflated distributions. The method primarily involves the model and notation, parameter estimation, and the mechanism for observing zero values.

We assume a continuous outcome variable Y, a mediator variable M, and an independent variable X. The outcome variable Y depends on the mediator M and the independent variable X through the following regression equation:

$$Y = \beta_0 + \beta_1 M + \beta_2 1(M > 0) + \beta_3 X + \beta_4 X 1(M > 0) + \beta_5 XM + \epsilon \tag{18.1}$$

where $1(\cdot)$ is an indicator function, the random error ϵ follows a normal distribution $N(0, \delta)$, and $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are regression coefficients. The mediator M represents the relative abundance (RA) of a microbial taxon.

To construct a mediation model, we also need to model the association between M and X. For the zero-inflated mediator M, we write its distribution into a two-part mixture distribution with the following density function:

$$f(m; \theta) = \begin{cases} G(\theta), & m = 0 \\ (1 - G(\theta))g(m; \theta), & m > 0 \end{cases} \tag{18.2}$$

where θ is the parameter vector associated with the zero-inflated distribution, $G(\cdot)$ is a mapping with $0 < G(\theta) < 1$ being the probability of M taking the value of 0, and $g(m; \theta)$ is the conditional probability density function of M given that M is positive. We use a zero-inflated Beta (ZIB) distribution for modeling the relative abundance of microbial taxa. We assume the parameter θ depends on X through the following equation:

$$T(\theta) = \nu_0 + \nu_1 X \tag{18.3}$$

where $T : \mathbb{R}^K \to \mathbb{R}^K$ is a known one-to-one (possibly nonlinear) transformation of the parameter vector θ, and ν_0 and ν_1 are parameter vectors interpreted as the intercept and slope, respectively. The above two equations together form our full mediation model.

Mechanism for Observing Zero Values In microbiome abundance data, observations below the limit of detection are set to zero. Consequently, the observed zero values include true zeros (i.e., the

actual absence of microbial taxa) and false zeros (i.e., the presence of microbial taxa that were not detected). We consider the following mechanism for observing zeros:

$$P(M^* = 0|M, L) = 1(ML < 1) \tag{18.4}$$

where L is the library size (sequencing depth), and ML is the sample absolute abundance (SAA) of the taxon. Under this mechanism, all SAA below 1 have an observed value of zero.

Mediation Effect and Direct Effect Under the potential outcomes framework, we define the natural indirect effect (NIE), natural direct effect (NDE), and controlled direct effect (CDE), where NIE is the mediation effect. The total effect of the exposure variable X is equal to the summation of NIE and NDE. For X changing from x_1 to x_2, the NIE, NDE, and CDE are defined as follows:

$$\text{NIE} = E(Y_{x_2 M_{x_2}} - Y_{x_2 M_{x_1}}) \tag{18.5}$$

$$\text{NDE} = E(Y_{x_2 M_{x_1}} - Y_{x_1 M_{x_1}}) \tag{18.6}$$

$$\text{CDE} = E(Y_{x_2 m} - Y_{x_1 m}) \tag{18.7}$$

where $Y_{x_2 M_{x_1}}$ is a counterfactual outcome. By plugging the regression equations into the above definitions, we obtain the following formulas:

$$\text{NIE} = (\beta_1 + \beta_5 x_2)(E(M_{x_2}) - E(M_{x_1})) + (\beta_2 + \beta_4 x_2)(E(1(M_{x_2} > 0)) - E(1(M_{x_1} > 0))) \tag{18.8}$$

$$\text{NDE} = \beta_3(x_2 - x_1) + \beta_4(x_2 - x_1)(1 - G(\theta_{x_1})) + \beta_5(x_2 - x_1)E(M_{x_1}) \tag{18.9}$$

Parameter Estimation We use maximum likelihood estimation (MLE) to estimate the parameters. For each subject i, the observed data can be denoted by the vector $(Y_i, R_i, M_i^*, L_i, X_i)$, where $R_i = 1(M_i^* > 0)$. The complete log-likelihood function is given by:

$$\ell = \sum_{i \in \text{group 1}} \ell_i^1 + \sum_{i \in \text{group 2}} \ell_i^2 \tag{18.10}$$

where the first group consists of subjects whose observed values are non-zero, and the second group consists of subjects whose observed values are zero.

18.2 Pros and Cons

18.2.1 Pros

1. **Ability to Handle Zero-Inflated Data**: The MedZIM method can decompose the mediation effect into two parts: one attributable to changes in positive relative abundance and the other to binary changes from zero to a non-zero state. This decomposition provides a more nuanced understanding and analysis of zero-inflated data (see Section 1, Abstract, Page 2). 2. **Handling False Zeros**: By using probabilistic models, MedZIM can address the issue of false zeros, which are data points that are observed as zero but are actually non-zero. This enhances the accuracy and reliability of the method (see Section 1, Abstract, Page 2).

Comparison with CCMM and SparseMCMM

18.2.2 Cons

1. **Complexity**: The implementation of this method involves complex probabilistic models and maximum likelihood estimation, which can increase the difficulty of implementation and interpretation (see Section 3, Paragraph 2, Page 6).

2. **Dependence on Assumptions**: Although the method addresses zero inflation and false zeros, it still relies on certain assumptions (such as the mechanism for observing zeros). The correctness of these assumptions can affect the performance of the model (see Section 6, Paragraph 2, Page 14).

3. **High Computational Resource Demand**: Due to the need to handle complex mixture distributions and zero-inflated data, MedZIM requires significant computational resources, which may pose performance bottlenecks in large-scale datasets (see Section 4, Paragraph 3, Page 8).

18.3 counterfactual assumption

The method described in the paper relies on the potential outcomes framework, which means it requires several standard counterfactual assumptions for causal inference. These assumptions include:

1. **No Unmeasured Confounders (Exchangeability)**: It is assumed that, given the covariates, there are no unmeasured confounders affecting both the exposure and the outcome. This means that all confounding factors that could influence the exposure and the outcome have been measured and controlled for (see Section 6, Paragraph 3, Page 14).

2. **Consistency**: It is assumed that the potential outcome under the observed exposure level for each individual is equal to the observed outcome. In other words, if an individual is observed under a certain exposure condition, their potential outcome is their actual outcome (see Section 6, Paragraph 3, Page 14).

3. **Positivity**: It is assumed that every individual has a non-zero probability of being assigned to each exposure level. This means that for all combinations of covariates, individuals could potentially receive different exposure conditions (see Section 6, Paragraph 3, Page 14).

18.4 Simulation

In the simulation part of the paper, the data generation method is as follows:

1. **Independent Variable** X: The independent variable X is binary and generated using a Bernoulli distribution Ber(0.5) such that the number of subjects is balanced between the two groups.

2. **Mediator Variable** M: The relative abundance (RA) data for the mediator variable M is generated using zero-inflated Beta (ZIB) distributions, mimicking the distribution in real data. The ZIB distribution generation is based on model equation (1) and equations (3-5).

3. **Library Size**: The library size is randomly picked from real study data, ranging from 31,607 to 911,652.

4. **Zero-Value Mechanism**: Zero values are generated using the limit of detection mechanism as per equation (6), where all sample absolute abundances (SAA) below 1 have an observed value of zero.

5. **Sample Size and Dimensions**: The sample size in the simulations is 100, with 100 random datasets generated.

19 CMMB, A compositional mediation model for a binary outcome to microbiome studies

19.1 Formula

In this article, we extend CMM to accommodate binary outcomes. The effect of a treatment on all the components of a compositional mediator is jointly estimated using the algebra in the simplex space.

Algebraic Operators in Simplex Space

First, we provide the definitions of the algebraic operators in the simplex space that appear in this article. For two compositions of k-components $g, f \in S^{k-1}$, the perturbation operator is defined as:

$$g \oplus f = \left(\frac{g_1 f_1}{\sum_{j=1}^{k} g_j f_j}, \ldots, \frac{g_k f_k}{\sum_{j=1}^{k} g_j f_j} \right)$$

the inverse of the perturbation operator is:

$$g \ominus f = \left(\frac{g_1 f_1^{-1}}{\sum_{j=1}^{k} g_j f_j^{-1}}, \ldots, \frac{g_k f_k^{-1}}{\sum_{j=1}^{k} g_j f_j^{-1}} \right)$$

the power transformation for a composition g by a scalar t is:

$$g^t = \left(\frac{g_1^t}{\sum_{j=1}^{k} g_j^t}, \ldots, \frac{g_k^t}{\sum_{j=1}^{k} g_j^t} \right)$$

and the norm for composition is:

$$\|g\| = (\log(g)^T N^{-1} \log(g))^{1/2}$$

where $\log(\cdot)$ represents the additive log-ratio (alr) transformation, and N^{-1} is the inverse of the $(k-1)\times(k-1)$ matrix $N = I_{k-1} + 1_{k-1}1_{k-1}^T$.

Compositional Mediation Model for Binary Outcomes

Suppose that we have n random samples from a population, where we observe an outcome Y_i, a compositional mediator M_i, a treatment T_i, and covariates X_i for $i = 1, \ldots, n$. We consider an expected causal effect of T_i on Y_i mediated through M_i, depicted in Figure 1. Then, a model for this mediation effect should take the compositional nature of M_i into account, as $M_i \in S^{k-1}$.

With the perturbation and power transformation operators, the proposed compositional mediation model for a binary outcome (CMM) is given by:

$$M_i = m_0 \oplus aT_i \oplus \bigoplus_{r=1}^{p} h_r X_{ri} \oplus U_{1i}$$

$$Y_i = 1\{c_0 + cT_i + b^T \log(M_i) + g^T X_i + U_{2i} > 0\}, \quad \text{subject to} \quad 1_k^T b = 0$$

where m_0 is a baseline composition, c_0 is a baseline measure for Y_i, a is a vector of composition parameters for a treatment, c are regression coefficients for the treatment, b are regression coefficients for the composition, $h_1, \ldots, h_p$ and g are nuisance parameters corresponding to X_i, U_{1i} and U_{2i} are disturbance terms for M_i and Y_i, respectively. We assume U_{1i} follows a logistic normal distribution with mean 0 and covariance Σ and U_{2i} follows a standard normal distribution. Model (1) formulates the effect of a treatment on a compositional mediator perturbed from the baseline composition, and Model (2) links treatment and a compositional mediator to a binary outcome after adjusting for pretreatment covariates while accounting for the compositional nature of M_i by imposing a zero-sum constraint.

Model Assumptions and Identification

As in most of the work on causal mediation analysis, estimators of the natural direct and indirect (or mediation) effects for the proposed method are defined under the causal assumptions: the stable unit treatment value assumption (SUTVA), the positivity assumption, and the no-unmeasured confounding assumption. Suppose that Models (1) and (2) are correctly specified. Then, under these assumptions, the direct effect $\phi(t)$ and the total indirect effect $\delta(t)$ are identifiable and given by:

$$\phi(s) = \mathbb{E}[Y_i(t, \log M_i(s)) - Y_i(t_0, \log M_i(s))|X_i = x] = \mathbb{E}\left[\Phi\left(\frac{ct + f(s, X_i)}{\sqrt{b^T\Sigma b + 1}}\right) - \Phi\left(\frac{ct_0 + f(s, X_i)}{\sqrt{b^T\Sigma b + 1}}\right)\right]$$

$$\delta(s) = \mathbb{E}[Y_i(s, \log M_i(t)) - Y_i(s, \log M_i(t_0))|X_i = x] = \mathbb{E}\left[\Phi\left(\frac{\log a^T bt + g(s, X_i)}{\sqrt{b^T\Sigma b + 1}}\right) - \Phi\left(\frac{\log a^T bt_0 + g(s, X_i)}{\sqrt{b^T\Sigma b + 1}}\right)\right]$$

where $\Phi(\cdot)$ is the standard normal distribution function, Σ is the covariance matrix of U_{1i}, and $f(s, X_i)$ and $g(s, X_i)$ are functions involving covariates X_i.

Estimation of Composition Parameters

To estimate the parameters in Model (1), we minimize the difference between observed and estimated compositions in S^{k-1}. Specifically, we solve the following optimization problem:

$$\hat{h} = \arg \min_{m_0, a, h_r \in S^{k-1}} \sum_{i=1}^{n} \|M_i - (m_0 \oplus aT_i \oplus \bigoplus_{r=1}^{p} h_r X_{ri})\|^2$$

where $\hat{h} = (\hat{m}_0, \hat{a}, \hat{h}_1, \ldots, \hat{h}_p)$.

Estimation of Regression Parameters

To estimate the parameters of the composition in Model (2), we use a log-contrast model. Let $g_i = 2y_i - 1$, $z_i = (1, t_i, \log(m_i)^T, x_i^T)^T$, $\beta = (c_0, c, b^T, g^T)^T$, and $q(g_i z_i^T \beta) = -\log \Phi(g_i z_i^T \beta)$. Then, the $L1$ penalized log-likelihood function for Model (2) is given by:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} q(g_i z_i^T \beta), \quad \text{subject to} \quad \|\beta\|_1 \leq \tau, \quad 1_k^T b = 0$$

where $\tau \geq 0$ is some constant. The solution of this optimization problem is equivalent to the solution of the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \|N^{1/2}(u - Z\beta)\|_2^2 + \lambda \|\beta\|_1, \quad 1_k^T b = 0$$

where N is an $n \times n$ diagonal matrix with its ith diagonal term $N_{ii} = \nu_i(g_i z_i^T \beta)[z_i^T \beta + \nu_i(g_i z_i^T \beta)]$, $\nu_i(\beta) = g_i \phi(g_i z_i^T \beta)/\Phi(g_i z_i^T \beta)$, $Z = (z_1, \ldots, z_n)^T$, $u = Z\beta_0 + N^{-1}\nu(\beta_0)$, and $\lambda \geq 0$ is a penalty term.

By transforming Z into $\tilde{Z} = Z(I_p - 1_k 1_k^T/k)$, the optimization problem can be written as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \|N^{1/2}(u - \tilde{Z}\beta)\|_2^2 + \lambda \|\beta\|_1, \quad 1_k^T b = 0$$

We propose a method that combines iteratively reweighted least squares (IRLS) and coordinate descent method of multipliers (CDMM). The algorithm steps are as follows:

- Construct the augmented Lagrangian:

$$L_\lambda = \frac{1}{2n} \|N^{1/2}(u - \tilde{Z}\beta)\|_2^2 + \lambda \|\beta\|_1 + \xi^T (1_k^T \beta) + \frac{\alpha}{2} (1_k^T \beta)^2$$

  where ξ is the Lagrange multiplier, and $\alpha > 0$ is the penalty parameter.

- Iterative updates:

$$\beta^{(l+1)} \leftarrow \arg\min_{\beta} L_\lambda(\beta, N^{(l)}, u^{(l)}, \xi^{(l)})$$

$$N^{(l+1)} \leftarrow \arg\min_{\beta} \sum_{i=1}^{n} q(g_i \tilde{z}_i^T \beta)$$

$$u^{(l+1)} \leftarrow \tilde{Z}\beta^{(l+1)} + N^{-1}\nu(\beta^{(l+1)})$$

$$\left(\frac{1}{\alpha}\right)^{(l+1)} \leftarrow \left(\frac{1}{\alpha}\right)^{(l)} + 1_k^T \beta$$

Debiasing Procedure and its Asymptotic Convergence

The solution $\hat{\beta}$ of Equation (5) is biased because of $L1$ penalization. To correct this bias, we adapt the debiasing procedure of Shi et al. (2016) and Lu et al. (2019). The proposed debiased estimator of $\hat{\beta}$ given N and u is:

$$\hat{\beta}_{db} = \hat{\beta} + \frac{1}{n} \tilde{H} \tilde{Z}^T N(u - \tilde{Z}\hat{\beta})$$

where $\tilde{H} = (I_p - 1_k 1_k^T/k)H$, and $H = (\hat{h}_1, \ldots, \hat{h}_p)^T$ is a solution of the following convex problem:

$$\hat{h}_j = \min_{h_j} h_j^T R h_j, \quad \text{subject to} \quad \|R h_j - (I_p - 1_k 1_k^T/k) e_j\|_1 \leq \gamma$$

where $j = 1, \ldots, p$, $R = \tilde{Z}^T N \tilde{Z}/n$, e_j is the vector with one at the jth position and zero everywhere else, and γ is some constant. Under some regularity conditions, the debiased estimator $\hat{\beta}_{db}$ converges to β as $n, p \to \infty$.

Hypothesis Test of Mediation Effect

The distribution of the total mediation effect $\delta(s)$ is unknown, so we propose a bootstrap approach to test the significance of an expected causal mediation effect:

$$H_0 : \delta(s) = 0 \quad \text{versus} \quad H_1 : \delta(s) \neq 0$$

To construct a sampling distribution of $\delta(s)$, we repeat the following steps B times: (i) randomly select n samples from the original n samples with replacement, and (ii) estimate $\delta_b(s)$. We use the 95% percentile confidence interval to test the significance of $\delta(s)$ in this study. Alternatively, we can estimate an approximate p-value for $\delta(s)$ utilizing the fact that any bootstrap replicate $\delta_b(s) - \delta(s)$ should have a distribution close to that of $\delta(s)$ when the null hypothesis is true.

19.2 Pros and Cons

19.2.1 Pros

1. **Handling Compositional Characteristics**: The method properly handles the compositional characteristics of microbiome data using algebraic operations in simplex space, avoiding undesirable consequences of neglecting this structure. This is highlighted on page 17, first paragraph .

2. **Handling Multiple Mediators**: The method is capable of handling multiple mediators simultaneously, making it suitable for high-dimensional data. This is mentioned on page 16, second paragraph .

3. **Sensitivity Analysis**: The method provides a sensitivity analysis approach for the no unmeasured confounding assumption, enhancing the robustness of the results. This is described on page 19, fourth paragraph.

19.2.2 Cons

1. **High Computational Complexity**: The method requires substantial computational resources, especially for high-dimensional data and sensitivity analysis. This is mentioned on page 21, second paragraph.

2. **Strong Assumption Dependence**: The method relies on assumptions such as no unmeasured confounding effects, which are generally not verifiable with observational data. This is highlighted on page 20, first paragraph.

3. **Not Ideal for Rare Outcomes**: The method is mainly developed for general outcome cases and may not be optimal for rare outcome situations. This is noted on page 20, second paragraph.

4. **Parameter Selection Sensitivity**: The debiasing procedure and penalty parameter selection may affect the stability and accuracy of the results. This is mentioned on page 20, first paragraph.

19.3 counterfactual assumption

1. **Stable Unit Treatment Value Assumption (SUTVA)**: This assumption posits that the treatment outcome for one unit does not affect the outcome of another unit, meaning there are no interference effects. Additionally, the potential outcomes for the same unit under different treatment conditions are independent. This is mentioned on page 18, second paragraph.

2. **Positivity Assumption**: This assumption requires that every combination of treatment and mediator variables has sufficient representation in the sample, meaning the probability of each treatment and mediator combination is greater than zero. This is also mentioned on page 18, second paragraph.

3. **No Unmeasured Confounding Assumption**: This assumption requires that there are no unmeasured confounders between the treatment and the outcome, and between the mediator and the outcome, meaning all potential confounding factors have been observed and controlled. This is highlighted on page 18, second paragraph.

19.4 Simulation

In the simulation part of the article, the data generation method includes: generating the treatment variable T_i from a Bernoulli distribution, the compositional disturbance term U_{1i} from a multivariate logistic normal distribution, and the outcome disturbance term U_{2i} from a standard normal distribution. The compositional mediator variable M_i is generated with $a = (20, 10, 5, 2, 1_{k-4})/(20, 10, 5, 2, 1_{k-4})1_k$, and the outcome variable Y_i is generated with $b = (0.5, -0.5, 0.5, -0.5, 0_{k-4})$. The baseline composition m_0 is set to $1_k/k$. The data dimensions are $k = 5, 25, 50$ and the sample size is $n = 50$. These steps are described in the second paragraph on page 19.

20 MarZIC

20.1 Formula

same as Medzim

20.2 Pros and Cons

20.2.1 Pros

20.2.2 Cons

20.3 counterfactual assumption

20.4 Simulation

21 MedFix, MedMix

21.1 Formula

22 Methodology

In this section, we describe two methods for high-dimensional mediator selection: MedFix and MedMix.

MedFix

For the model:

$$Y = X\alpha + M\gamma + Z\beta + \epsilon$$

$$M_j = XA_j + ZB_j + \epsilon_j, \quad j = 1, \ldots, p$$

We apply sparse regression techniques such as adaptive lasso to fit the mediator models and outcome model separately. The outcome model has heterogeneous predictors (M is continuous, Z is discrete), so we introduce an additional tuning parameter to adjust for this heterogeneity. We estimate the parameters by minimizing the following objective function:

$$(\hat{\alpha}_{\text{fix}}, \hat{\gamma}_{\text{fix}}, \hat{\beta}) = \arg\min_{\alpha,\gamma,\beta} \left\{ \frac{1}{2}\|Y - X\alpha - M\gamma - Z\beta\|_2^2 + \rho(\gamma, \beta) \right\}$$

where

$$\rho(\gamma, \beta) = \lambda(1-\eta)\sum_{j=1}^{p} w_j^{(\gamma)}|\gamma_j| + \lambda\eta\sum_{k=1}^{q} w_k^{(\beta)}|\beta_k|$$

Here, $\lambda > 0$ is the overall penalty tuning parameter, $\eta \in (0, 1)$ adjusts the regularization levels on the two data types, and the weight vectors $w^{(\gamma)}$ and $w^{(\beta)}$ are normalized to sum to p and q, respectively. We refer to this solution as the MedFix method.

Causal Effect Testing

We perform statistical tests for the natural indirect effect (NIE), natural direct effect (NDE), total effect (TE), and individual mediator effects (NIE-j). Using the oracle property of adaptive lasso, we can reformulate these tests as tests on the regression coefficients. For the MedFix method, the p-value for NIE-j is obtained by a z-test:

$$P_{\gamma_j} = 2[1 - \Phi(|\hat{\gamma}_j|/s_j)] \quad \text{if } \hat{\gamma}_j \neq 0$$

$$P_{\gamma_j} = 1 \quad \text{otherwise}$$

where Φ is the cumulative distribution function of the standard normal distribution. The p-value for Bj is obtained from a χ^2 likelihood ratio test. A stepdown procedure is applied to control the false discovery proportion (FDP) in mediator selection.

PVM Estimation

We estimate VDE, VIE, and VTE in (3) using the following formulas:

$$\hat{VDE}_{\text{fix}} = \frac{1}{n}\|Z\hat{\beta}\|^2$$

$$\hat{VIE}_{\text{fix}} = \frac{1}{n}\|Z(\hat{B}\hat{\gamma})\|^2$$

$$\hat{VTE}_{\text{fix}} = \frac{1}{n}\|Z(\hat{B}\hat{\gamma} + \hat{\beta})\|^2$$

$$\hat{PVM}_{\text{fix}} = \frac{\hat{VIE}_{\text{fix}}}{\hat{VTE}_{\text{fix}}}$$

MedMix Method

Parameter Estimation

For the model:

$$Y = X\alpha + M\gamma + u + \epsilon$$
$$M_j = XA_j + v_j + \epsilon_j, \quad j = 1, \ldots, p$$

We model the aggregate effects of Zβ and ZBj as random effects, thereby reducing the number of candidate parameters. The resulting linear mixed models are:

$$Y = X\alpha + M\gamma + u + \epsilon$$
$$M_j = XA_j + v_j + \epsilon_j, \quad j = 1, \ldots, p$$

where $\epsilon \sim N(0, \sigma^2 I_n)$, $\epsilon_j \sim N(0, \sigma_j^2 I_n)$, u and v-j are random effect vectors of length n, and independent of the noise vectors. We estimate the parameters by minimizing the following penalized negative log-likelihood:

$$Q_{\lambda,w}(\alpha, \gamma, \tau, \sigma^2) = \frac{1}{2}(Y - X\alpha - M\gamma)^T V^{-1}(Y - X\alpha - M\gamma) + \frac{1}{2}\log|V| + \lambda w(\gamma)$$

where $V = \tau K(Z) + \sigma^2 I_n$. We propose a novel variable selection algorithm to solve this non-convex problem and term this method as MedMix.

Causal Effect Testing

For MedMix, the causal effect testing is similar to MedFix, with the following differences. For testing individual mediator effects, we replace the null hypothesis $H0, Bj = 0$ with $H0, \gamma_j = 0$, and the p-value P_{γ_j} is calculated using a SKAT-like score test. The test statistic for γ in the outcome model is:

$$Q_\tau = (Y - X\hat{\alpha} - M\hat{\gamma})^T K(Z)(Y - X\hat{\alpha} - M\hat{\gamma})$$
$$Q_j = (M_j - X\hat{A}_j)^T K(Z)(M_j - X\hat{A}_j)$$

The p-value for $H0, j$ is $P_{\text{med},j} = \max(P_{\gamma_j}, P_{Bj})$. For testing NDE, we propose to test $H0, \tau = 0$ with its p-value given by a similar score test. Testing TE can be done separately using the model $Y = X\alpha + \tilde{u} + \epsilon$, where $\tilde{u} \sim N(0, \tau K(Z))$ and $\epsilon \sim N(0, \tilde{\sigma}^2 I_n)$. A simple score test is used:

$$Q_\tau = (Y - X\hat{\alpha})^T K(Z)(Y - X\hat{\alpha})$$

PVM Estimation

Let $\Psi = (\hat{v}_1, \ldots, \hat{v}_p)$ be the n × p matrix of estimated random effects. The MedMix-based estimates of causal effects are:

$$\hat{VDE}_{\text{mix}} = \frac{1}{n}\|\hat{u}\|^2$$
$$\hat{VIE}_{\text{mix}} = \frac{1}{n}\|\Psi\hat{\gamma}\|^2$$
$$\hat{VTE}_{\text{mix}} = \frac{1}{n}\|\Psi\hat{\gamma} + \hat{u}\|^2$$
$$\hat{PVM}_{\text{mix}} = \frac{\hat{VIE}_{\text{mix}}}{\hat{VTE}_{\text{mix}}}$$

22.1 Pros and Cons

22.1.1 Pros

MedFix Method

1. *Simplicity and Intuitiveness*: MedFix is based on a fixed effects regression framework, using adaptive lasso for parameter estimation, making the method relatively simple and straightforward (page 435, paragraph 1). 2. *Strong Theoretical Foundation*: The adaptive lasso method possesses the oracle property, maintaining good variable selection performance in high-dimensional settings (page 437, paragraph 2).

MedMix Method

1. *Robustness*: MedMix is based on high-dimensional linear mixed models, making it more robust against model misspecification (page 437, paragraph 2). 2. *Good False Discovery Rate Control*: MedMix is better at controlling the false discovery rate in high-dimensional settings (page 446, paragraph 4). 3. *Suitable for Complex Models*: By modeling aggregate effects as random effects, MedMix reduces the number of candidate parameters, making it suitable for more complex models (page 438, paragraph 2).

22.1.2 Cons

MedFix Method 1. *Sensitivity to Model Assumptions*: MedFix is sensitive to model assumptions, and its performance significantly decreases when the model assumptions deviate (page 442, paragraph 3). 2. *Potential for High False Positives*: In high-dimensional settings, MedFix may select too many false positive mediators, increasing errors (page 446, paragraph 3). 3. *High Computational Complexity*: The need to apply adaptive lasso regression to each mediator results in high computational complexity (page 437, paragraph 1).

MedMix Method 1. *Complexity*: MedMix requires more complex parameter estimation and variable selection algorithms, making the method relatively complicated (page 438, paragraph 2). 2. *Need for Initial Estimates*: MedMix requires initial estimates, which may affect the results (page 447, paragraph 3). 3. *High Computational Resource Demand*: Due to the estimation of random effects, MedMix demands high computational resources (page 439, paragraph 3).

22.2 counterfactual assumption

Specifically, the assumptions include: 1. *Ignorability Assumption*: It is assumed that, given the covariates, the exposure and mediator are independent of the outcome distribution. 2. *Stable Unit Treatment Value Assumption (SUTVA)*: It is assumed that each unit's outcome is only affected by its own exposure and mediator, not by those of other units. 3. *Counterfactual Consistency Assumption*: It is assumed that each unit's outcomes under different exposure and mediator conditions are comparable.

22.3 Simulation

In the simulation part of the paper, the data generation method is based on the analysis of a mouse f2 dataset. The simulation model combines a linear fixed effects model and a linear mixed effects model, controlled by a simulation parameter θ ranging from [0, 1]. When $\theta = 0$, the fixed effects model is the true model; when $\theta = 1$, the mixed effects model is the true model. The simulated data includes genotype, preprocessed islet gene expression, and gender data. The simulated outcomes Y and mediators M have dimensions of approximately $n = 491$, including 15 true mediators, 15 fake mediators, and around 11,000 null candidate mediators.

23 SparseMCMM-HD, A microbial causal mediation analytic tool for health disparity and applications in body mass index

23.1 Formula

In this paper, we extend SparseMCMM to a non-manipulable exposure setting, propose a microbial causal mediation framework for health disparity study, and denote it as SparseMCMM-HD.

Instead, one can interpret the causality of health inequality by the hypothesized intervention efect on the manipulable mediating variable. Tus, in SparseMCMM-HD, we aim to quantify the overall health inequality on the outcome (called overall disparity), the health inequality efect that would be eliminated by equalizing microbiome profles across ethnic or regional groups (called manipulable disparity), and the healthy inequality efect that would remain even after microbiome profles across ethnic or regional groups were equalized (called residual disparity).

ausal Mediation Model Suppose there are I subjects from two categories of a non-manipulable exposure group (e.g., ethnicity or region), J taxa, and K covariates. Subscripts i, j, and k indicate a subject, a taxon, and a covariate, respectively. For the ith subject, let $R_i = 1$ or 0 indicate the reference or comparison group, let $\mathbf{M}_i = (M_{i1}, \ldots, M_{iJ})^T$ be the microbiome relative abundance vector with the constraint $\sum_{j=1}^{J} M_{ij} = 1$, and let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})^T$ represent the covariates, and let Y_i be a continuous outcome of interest.

To statistically describe the causal relationships shown in Figure 1, following our previous work, we use the linear log-contrast model to regress the continuous outcome on the non-manipulable exposure, microbiome compositions, and interactions between the non-manipulable exposure and microbiome compositions, while adjusting the confounding covariates:

$$Y_i = \alpha_0 + \alpha_R R_i + \alpha_X^T \mathbf{X}_i + \alpha_M^T \log(\mathbf{M}_i) + \alpha_C^T (\log(\mathbf{M}_i) \cdot R_i) + \epsilon_i \tag{23.1}$$

where α_0 is the intercept, α_R is the coefficient of the non-manipulable exposure, $\alpha_X = (\alpha_{X1}, \ldots, \alpha_{XK})^T$, $\alpha_M = (\alpha_{M1}, \ldots, \alpha_{MJ})^T$, and $\alpha_C = (\alpha_{C1}, \ldots, \alpha_{CJ})^T$ are the vectors of coefficients of covariates, microbiome compositions, and interactions between the non-manipulable exposure and microbiome compositions, respectively. Due to the compositionality of the microbiome data $\sum_{j=1}^{J} M_{ij} = 1$, α_M and α_C are additionally subject to $\alpha_M^T 1 = 0$, and $\alpha_C^T 1 = 0$. $\epsilon_i \sim N(0, \sigma^2)$ is the error term.

On the other hand, the Dirichlet regression is used to model the microbial relative abundance as a function of the non-manipulable exposure and covariates:

$$\mathbf{M}_i | (R_i, \mathbf{X}_i) \sim \text{Dirichlet}(\gamma_1(R_i, \mathbf{X}_i), \ldots, \gamma_J(R_i, \mathbf{X}_i)) \tag{23.2}$$

Specifically, we assume that the microbial relative means are linked with the non-manipulable exposure and covariates $(R_i, \mathbf{X}_i)$ in the generalized linear model fashion with a log link:

$$\log(\gamma_j(R_i, \mathbf{X}_i)) = \beta_{0j} + \beta_{Rj} R_i + \beta_X^T \mathbf{X}_i \tag{23.3}$$

Definition of Disparity Measures in the Counterfactual Framework As discussed in the "Background" section, we propose to conceptualize an overall disparity measure (ODM) on the outcome that can be decomposed into manipulable disparity measure (MDM) and residual disparity measure (RDM). MDM represents the portion of disparity that would be eliminated by equalizing microbiome profiles between comparison and reference groups, and RDM represents the portion that would remain even after microbiome profiles between comparison and reference groups were equalized. With the counterfactual notation, mathematically we have:

$$\text{ODM} = \text{MDM} + \text{RDM} \tag{23.4}$$

MDM, RDM, and ODM can be further expressed as follows:

$$\text{MDM} = E\left[E[Y_{M_{x(1)}} | R = 1, x]\right] - E\left[E[Y_{M_{x(0)}} | R = 1, x]\right] \tag{23.5}$$

$$\text{RDM} = E\left[E[Y_{M_{x(0)}} | R = 1, x] - E[Y_{M_{x(0)}} | R = 0, x]\right] \tag{23.6}$$

$$\text{MDM} = \sum_{j=1}^{J} (\alpha_{Mj} + \alpha_{Cj}) \left(E[\log(M_j) | R = 1, x] - E[\log(M_j) | R = 0, x]\right) \tag{23.7}$$

$$\text{RDM} = \alpha_R + \alpha_C^T E[\log(M) | R = 0, x] \tag{23.8}$$

Parameter Estimation Analogous to SparseMCMM, we employ a two-step procedure to estimate the regression parameters in models (1)–(2) to obtain the estimated RDM, MDM, and MDM for each taxon, and ODM. Furthermore, SparseMCMM-HD has the full capability to perform variable selection to select the signature causal microbes that play mediating roles in the disparity of the continuous outcome with regularization strategies. Specifically, L1 norm and group-lasso penalties are incorporated for variable selection. To account for the biases introduced by the regularization techniques employed, we further implement splitting strategy, which can handle arbitrary penalties and provide asymptotically validated inference. We also incorporate this splitting strategy in the SparseMCMM package to refine its estimation procedure.

Hypothesis Tests for Manipulable Disparity Similarly, we employ the hypothesis tests for mediation effects in SparseMCMM to examine whether the microbiome has any mediation effect on the disparity in the outcome, at the community and taxon levels, respectively. Specifically, regarding the null hypothesis of no manipulable disparity $H_0 : \text{MDM} = 0$, the first test statistic is defined as $\text{OMD} = \text{M}\hat{\text{D}}\text{M}$, the estimator of the manipulable disparity. OMD examines whether or not the whole microbiome plays a mediating role in health disparities. Meanwhile, we consider another null hypothesis, $H_0 : \text{MDM}_j = 0, \forall j \in \{1, \ldots, J\}$ and define the second test statistic as $\text{CMD} = \sum_{j=1}^{J} \text{M}\hat{\text{D}}\text{M}_j^2$, the summation of squared estimators of individual mediation effects across all taxa. CMD examines whether or not at least one taxon mediates the health disparities. Permutation procedure is employed to assess the significance of these two test statistics. This provides a mechanism to check whether the microbiome has any impact on health disparity that could be potentially eliminated through the microbiome.

23.2 Pros and Cons

23.2.1 Pros

1. **Rigor and Validity of the Model**: SparseMCMM_HD is a rigorous and validated causal mediation analysis framework that can effectively identify mediating microbes under the high-dimensional and sparse structure of microbiome data. The model uses regularization techniques to handle microbiome features, effectively selecting signature causal microbes involved in health disparities [**?** , p.3, para.2].

2. **Capability to Handle Complex Data**: This framework can handle high-dimensional, sparse, and compositional microbiome data, which is difficult to achieve with traditional mediation analysis methods [**?** , p.2, para.2].

3. **Potential to Address Health Disparities**: By quantifying manipulable disparities, SparseMCMM_HD provides a viable path to reducing health disparities through microbiome modulation, demonstrating significant potential in reducing BMI disparities between different ethnicities or regions [**?** , p.11, para.1].

23.2.2 Cons

1. **Limitation on Dynamic Features**: The current SparseMCMM_HD framework deals with microbiome data at a fixed time point, which does not fully accommodate the dynamic nature of the microbiome, potentially limiting the model's accuracy in practical applications [**?** , p.11, para.2].

2. **Limitation in Handling Categorical Outcomes**: The framework primarily addresses continuous outcomes. It needs further development to handle multiple binary or categorical outcomes, limiting its application in some health disparity studies [**?** , p.11, para.2].

3. **Limitation of the Data Splitting Strategy**: Due to the inference-prediction tradeoff, the data splitting strategy is constrained, particularly with smaller sample sizes, which may affect the robustness and accuracy of the estimates [**?** , p.11, para.2].

4. **Challenges in Controlling for Confounding Factors**: Although propensity score matching (PSM) is used to control for confounding factors, PSM may not adequately account for all latent variables influencing health disparities analysis in limited data situations, impacting the accuracy of causal inference [**?** , p.11, para.2].

23.3 counterfactual assumption

1. **No Unmeasured Confounding for the Effect of Non-manipulable Exposure on the Outcome**: It is assumed that the effect of the non-manipulable exposure R on the outcome Y is unconfounded conditional on all covariates X, i.e., $Y \perp R \mid X$ [**?** , p.4, para.2].

2. **No Unmeasured Confounding for the Effect of the Mediator on the Outcome**: It is assumed that the effect of the mediator M on the outcome Y is unconfounded conditional on the non-manipulable exposure R and all covariates X, i.e., $Y \perp M \mid R, X$ [**?** , p.4, para.2].

23.4 Simulation

In the simulation section of the paper, microbial relative abundance data is generated from a Dirichlet distribution to evaluate the performance of the SparseMCMM_HD method. The outcome variable is generated based on the assumed causal relationship model. The dimensions of the data include I subjects, J taxa, and K covariates, where I represents the number of subjects, J represents the number of taxa, and K represents the number of covariates.

24 DACT

24.1 Formula

Assumptions and Regression Models Let A denote the exposure, Y a continuous outcome, M a continuous mediator, and X additional covariates to adjust for confounding. Baron and Kenny (1986) proposed the following linear structural equation models:

$$Y = \beta_0 + \beta_A A + \beta_M M + \beta_X^T X + \epsilon_Y \tag{24.1}$$

$$M = \gamma_0 + \gamma_A A + \gamma_X^T X + \epsilon_M \tag{24.2}$$

where ϵ_Y and ϵ_M are error terms with mean zeros and constant variances. For a binary and rare outcome Y, the following logistic model can be used:

$$\text{logit}(P(Y = 1|A, M, X)) = \beta_0 + \beta_A A + \beta_M M + \beta_X^T X \tag{24.3}$$

The NIE (natural indirect effect) measures the mediation effect of an exposure on an outcome through a mediator and is defined as $\beta\gamma$.

Divide-Aggregate Composite-Null Test (DACT) The DACT leverages the whole genome DNA methylation data into the construction of the test statistic. For each DNA methylation CpG site, fit the outcome and mediator regression models to obtain p-values p_β for testing $\beta = 0$ and p_γ for testing $\gamma = 0$.

Based on the composite null hypothesis $H_0 : \beta\gamma = 0$, the testing is divided into three cases:

1. Case 1: $\beta \neq 0$, $\gamma = 0$, use p_γ.
2. Case 2: $\beta = 0$, $\gamma \neq 0$, use p_β.
3. Case 3: $\beta = 0$, $\gamma = 0$, use $(\text{MaxP})^2$.

To improve the detection of mediation effects in each case, we define the case-specific p-values as follows: - Case 1: Test for the exposure-mediator effect being zero ($\gamma = 0$), using p_γ. - Case 2: Test for the mediator-outcome effect being zero ($\beta = 0$), using p_β. - Case 3: Test for both the exposure-mediator and mediator-outcome effects being zero ($\beta = 0$ and $\gamma = 0$), using the squared p-value from the joint significance test $(\text{MaxP})^2$.

The composite p-value is then calculated as:

$$\text{DACT}_j = w_1 p_{\gamma,j} + w_2 p_{\beta,j} + w_3 (\text{MaxP}_j)^2 \tag{24.4}$$

where w_1, w_2, and w_3 are weights obtained by estimating the relative proportions of the three null cases.

Calibration of Empirical Null Distribution Use Efron's empirical null framework to calibrate the p-values of the DACT statistics:

$$Z_{\text{DACT}_j} = \Phi^{-1}(1 - \text{DACT}_j) \tag{24.5}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. The marginal probability density function $f(z)$ of Z_{DACT_j} is given by:

24.2 Pros and Cons

24.2.1 Pros

1. **Enhanced Power**: The DACT method has higher power in detecting mediation effects compared to traditional Sobel and MaxP tests (Section 3, Construction of DACT). 2. **Control of Type I Error Rate**: Simulation studies show that the DACT method can effectively control Type I error rates in finite samples, whereas Sobel and MaxP tests tend to be conservative in some cases (Section 3, Construction of DACT). 3. **Adaptation to Large-Scale Genomic Data**: By leveraging genome-wide data, the DACT method significantly improves the ability to detect mediation effects, which is crucial in large-scale genomic epigenetic studies (Section 3, Construction of DACT).

24.2.2 Cons

24.3 counterfactual assumption

(i) There are no unmeasured exposure–outcome confounders given X; (ii) There are no unmeasured mediator-outcome confounders given (X,A); (iii) There are no unmeasured exposure-mediator confounders given X; (iv) There is no effect of exposure that confounds the mediator-outcome relationship; (v) There is no exposure and mediator interaction on the outcome.

24.4 Simulation

25 MiMed

25.1 Formula

In this paper, the MiMed platform employs several causal mediation analysis methods, including the Sobel test, Preacher–Hayes approach, DACT method, and Imai method. These methods assess mediation effects using different statistical models and assumptions.

1. Sobel Test

The Sobel test evaluates the mediation effect using the following regression models:

$$M_i = a_0 + a_1 T_i + e_i \tag{25.1}$$

$$Y_i = b_0 + b_1 M_i + b_2 T_i + t_i \tag{25.2}$$

where T_i is the treatment variable, M_i is the mediator (e.g., an alpha-diversity index or a microbial taxon), Y_i is the health or disease outcome variable, a_0 and b_0 are intercepts, a_1, b_1, and b_2 are slopes, and e_i and t_i are independently distributed random errors.

The null and alternative hypotheses of the Sobel test are:

$$H_0 : a_1 b_1 = 0 \tag{25.3}$$

$$H_1 : a_1 b_1 \neq 0 \tag{25.4}$$

The Sobel test conducts significance testing using a parametric approach assuming that e_i and t_i are normally distributed.

2. Preacher–Hayes Approach

The Preacher–Hayes approach conducts significance testing non-parametrically using a bootstrap method without assuming normality.

3. Divide-Aggregate Composite-null Test (DACT)

The DACT method is also a parametric approach but considers the null hypothesis as a composite hypothesis to improve statistical power:

$$H_0 : a_1 b_1 = 0 \tag{25.5}$$

The composite null hypothesis includes the following three scenarios:

1. $a_1 = 0$ and $b_1 \neq 0$
2. $a_1 \neq 0$ and $b_1 = 0$
3. $a_1 = 0$ and $b_1 = 0$

4. Imai Method

The Imai method is based on the potential outcomes framework of causal inference, defining the unit-level total treatment effect, direct effect, and indirect effect as follows:

$$\delta_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \tag{25.6}$$

$$\phi_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \tag{25.7}$$

$$\zeta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \tag{25.8}$$

The overall average direct effect (ADE) is:

$$ADE = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i(0) + \frac{1}{n}\sum_{i=1}^{n}\phi_i(1)\right) \tag{25.9}$$

The overall average causal mediation effect (ACME) is:

$$ACME = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\zeta_i(0) + \frac{1}{n}\sum_{i=1}^{n}\zeta_i(1)\right) \tag{25.10}$$

The Imai method also allows for the consideration of interaction effects between the treatment and the mediator, modeled as follows:

$$Y_i = c_0 + c_1 T_i + c_2 M_i + c_3 T_i M_i + \epsilon_i \tag{25.11}$$

where $T_i M_i$ is the interaction term, c_0, c_1, c_2, and c_3 are regression coefficients, and ϵ_i is an independently distributed random error.

Using the above models, the Imai method provides interval estimation for total effect, direct effect, and indirect effect, and conducts significance testing using a bootstrap method.

25.2 Pros and Cons

25.2.1 Pros

25.2.2 Cons

25.3 counterfactual assumption

25.4 Simulation

26 Estimation of Mediation Effect on Zero-Inflated Microbiome Mediators

26.1 Formula

For mediation analysis, the interest is usually focused on separating the indirect pathway from the total causal effect between exposure and outcome. To better estimate these effects, we introduce counterfactual variables and use zero-inflated and zero-hurdle models to handle microbiome data with excessive zeros.

Zero-Inflated and Zero-Hurdle Models

Zero-inflated (ZI) models were first introduced by Cohen and widely accepted after Mullahy and Lambert. ZI models are mixtures of a discrete distribution and a zero point mass. Structural zeros are distinguished from count data, usually assumed to follow a Poisson or negative binomial (NB) distribution. The density function for ZI models can be defined as:

$$f_{ZI}(m_i) = \begin{cases} \phi_i, & \text{for } m_i = 0 \\ (1-\phi_i) g(m_i), & \text{for } m_i > 0 \end{cases}$$

On the other hand, zero-hurdle (ZH) models process the data in two stages to account for the excessive number of zeros. The first part is a binary model to determine whether the outcome is zero or a positive value. Logistic regression models are usually used for the first part to incorporate the effects of the covariates on the probability of an observation being zero. For the second part, distributions truncated at zero are used, conditioning on all the nonzero count outcomes. Zero-truncated regression models are then applied to incorporate the covariates effects on the nonzero distribution.

Inverse Probability Weighting Two-Part Model

To incorporate a variable with excessive zeros as the mediator, we decompose the mediation effect of the microbiome into two components inherent in the zero-inflated distributions: one attributed by a zero part ($M = 0$) and one attributed by a count part ($M > 0$). To simultaneously model the zero part and the count part of the mediator, we developed a weighting-based approach in which the estimation of exposure-covariate interactions and a separate averaging step can be avoided. Following Lange et al.'s work, we propose a semiparametric approach for estimating the direct and indirect effects while avoiding specification of the outcome model.

In this weighting-based approach, the estimation of exposure-covariate interactions and a separate averaging step are also avoided. Furthermore, the IPW approach is less computationally intensive and easier to formulate and implement. The weighting-based approach estimates the expectation of $E[Y(a, M(a^*))]$ as:

$$E[Y(a, M(a^*))] = E\left[Y \cdot I(A = a) \frac{f_{A|X}(a|x) \cdot f_{M|A,X}(m|a^*, x)}{f_{M|A,X}(m|a, x)} \right]$$

where the zero-inflated function is defined as:

$$f_{M|A,X}(m|a,x) = h(m^+|m_0,a,x) \cdot g(m_0|a,x)$$

Estimation of the Stabilized Direct and Indirect Effects

One of the major interests in mediation analysis lies in estimating the natural direct effect (NDE), natural indirect effect (NIE), and total effect (TE) of the counterfactual framework. Because we used the stabilized version of the weights, we named the targeted estimands as SNDE, SNIE, and STE.

We weight the outcomes by the inverse probability of each individual's exposure status and mediator levels by fitting a weighted logistic regression model.

26.2 Pros and Cons

26.2.1 Pros

1. **Addressing Zero-Inflation Issue**: The proposed method addresses the issue of zero inflation in microbiome data, which is not manageable by standard mediation analysis models. By decomposing zero-inflated distributions into zero part and count part, the method provides more accurate estimation of mediation effects (Section 2.1, Page 5).

2. The weighting-based approach avoids the specification of outcome models and exposure-covariate interactions, making the method more robust when dealing with complex data relationships (Section 2.2, Page 7).

26.2.2 Cons

26.3 counterfactual assumption

1. **Sequential Ignorability Assumption**: It is assumed that there is no unmeasured confounding of the exposure-outcome relationship, the exposure-mediator relationship, or the mediator-outcome relationship. Mathematically, it is expressed as:

$$Y(a, M(a)) \perp A|X, \quad M(a) \perp A|X, \quad Y(a,m) \perp M|(A,X)$$

(Section 2, Page 4)

2. **Consistency Assumption**: It is assumed that the counterfactuals take the observed values when the risk factor or treatment and mediator are actively set to the values they would have had. Mathematically, it is expressed as:

$$P(Y(A,M) = Y) = 1 \quad \text{and} \quad P(M(A) = M) = 1$$

(Section 2, Page 4)

3. **Positivity Assumption**: It is assumed that all levels of exposure and mediator have a nonzero probability for any values of the confounders. Mathematically, it is expressed as:

$$P(A = a|X = x) > 0 \quad \forall a, c \quad \text{and} \quad P(M = m|X = x, A = a) > 0 \quad \forall a, x, m$$

(Section 2, Page 4)

4. **Identification of Natural Effects Assumption**: In order to identify natural effects, it is assumed that the potential outcome $Y(a,m)$ is independent of the potential mediator $M(a^*)$ whenever a and a^* are different. Mathematically, it is expressed as:

$$Y(a,m) \perp M(a^*)|X \quad \text{for any} \quad m \quad \text{and} \quad a \neq a^*$$

(Section 2, Page 4)

26.4 Simulation

In the simulation section, the data generation method is as follows:

The dataset consists of 500 independent subjects, each with a binary exposure variable A, two confounders $X1$ and $X2$, and a binary outcome variable Y. The exposure variable A follows a binary distribution with an approximately 1:1 proportion. The confounder $X1$ follows a uniform distribution, $U \sim [10, 80]$, and the confounder $X2$ follows a binary distribution with a 50% proportion. The mediator variable M is generated using the following equations:

1. Exposure equation:
$$\text{logit}(A|X1, X2) = 0.4 - 0.005X1 + 0.05X2$$
2. Mediator equation (zero part):
$$\text{logit}(M_{m=0}|A, X1, X2) = -0.708A - 0.01X1 + 0.5X2$$
3. Mediator equation (count part):
$$\log(M_{m>0}|A, X1, X2) = 0.5 + \gamma^T A - 0.01X1 + 0.5X2$$
4. Outcome equation:
$$\text{logit}(Y|A, M, X1, X2) = -3 - 5A + M + 0.1X1 + X2$$

27 multimedia

27.1 Formula

Counterfactual Analysis

After using the `estimate` function to fit models to the observed data, we can reason about potential outcomes under different treatment regimes. This allows us to clarify the relative importance of direct and indirect pathways. For example, to estimate a direct effect (T $\rightarrow$ Y), we can block effects that travel along the indirect path (T $\rightarrow$ M $\rightarrow$ Y) and measure the changes to the response that persist.

The formula for direct effects is:

$$\zeta(t') = E\{Y_i(1, M_i(t')) - Y_i(0, M_i(t'))\}$$

The formula for indirect effects is:

$$\delta(t) = E\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\}$$

To increase modeling transparency, `multimedia` includes functions for interacting with and altering fitted models. Direct and indirect effects can be visualized within the context of the original data. This can serve as a sanity check and guide further model refinements.

Statistical Inference

The `multimedia` package offers bootstrap and synthetic null hypothesis testing approaches for quantifying uncertainty in estimates of mediation effects. For bootstrap in the mediation analysis context, we refit the mediation and outcome models to bootstrap resampled versions of the data and compute summary statistics (e.g., direct effect estimates) on each bootstrap sample to obtain the bootstrap distribution. The synthetic null hypothesis testing approach simulates synthetic data from an assumed null distribution to generate negative controls and calibrate a selection rule with false discovery rate control.

27.2 Pros and Cons

27.2.1 Pros

27.2.2 Cons

27.3 counterfactual assumption

27.4 Simulation

28 MicroBVS

28.1 Method

The authors proposed a statistical framework to analyze how the microbiome mediated the effect of a treatment on a health outcome via Bayesian method, called MicroBVS. The analysis focused on three components: the binary treatment or exposure (T), the compositional microbial mediators (M) and the continuous health outcome (Y). The authors developed a Bayesian joint model and defined the direct effect and indirect effect based on the model.

For the Treatment-Mediator pathway the authors apply Dirichlet-multinomial model, and they chose linear regression framework to model the Mediator-Outcome pathway. To link these two models together, they defined balance, a function of microbiome relative abundance, as the predictor of the linear regression model. In order to give the definition of balance, we need to first define partitions. For each microbial taxon k, the relative abundances $\boldsymbol{\psi}$ can be divided into two partitions ψ_{k+} and ψ_{k-}. Given a vector of relative abundances $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_J)$, $J-1$ sequential binary partitions can be generated. Specifically, the first partition is defined as $\psi_{1+} = \{\psi_1\}$ and $\psi_{1-} = \{\psi_2, \ldots, \psi_J\}$, the second partition is defined as $\psi_{2+} = \{\psi_2\}$ and $\psi_{2-} = \{\psi_3, \ldots, \psi_J\}$, and so on until $\psi_{J-1,+} = \{\psi_{J-1}\}$ and $\psi_{J-1,-} = \{\psi_J\}$. The elements of $\boldsymbol{\eta}_k$ take on values of $1, -1$, or 0 indicating the taxa positions in $\boldsymbol{\psi}$. 1 implies that the corresponding ψ_j belongs to partition ψ_{k+}, -1 that it belongs to partition ψ_{k-}, and 0 implies it is not in either partition. Therefore, the balance for a partition is defined as

$$B\left(\boldsymbol{\eta}_k, \boldsymbol{\psi}\right) = \sqrt{\frac{|\psi_{k+}|\,|\psi_{k-}|}{|\psi_{k+}| + |\psi_{k-}|}} \log\left(\frac{g\left(\psi_{k+}\right)}{g\left(\psi_{k-}\right)}\right), \tag{28.1}$$

where $|\cdot|$ indicates the dimension of the partition and $g(\cdot)$ the geometric mean.

The authors assume the microbiome relative abundances $\boldsymbol{\psi}_i$ for each subject i follow a Dirichlet-multinomial distribution to accommodate the overdispersion of microbiome data and later on facilitate the Markov chain Monte Carlo (MCMC) computation. Within this assumption, $\boldsymbol{z}_i \sim \text{Multinomial}\left(\dot{z}_i \mid \boldsymbol{\psi}_i\right)$, such that $\dot{z}_i = \sum_{j=1}^{J} z_{ij}$, and conjugate priors $\psi_i \sim \text{Dirichlet}\left(\boldsymbol{\gamma}_i\right)$, where $\boldsymbol{\gamma}_i$ is a J-dimensional vector of concentration parameters. The treatment model is a log-linear regression relating the relative abundances with the treatment.

$$log\left(\gamma_{ij}\right) = \alpha_j + \phi_j T_i + \sum_{p=1}^{P} \theta_{jp} x_{ip}, \tag{28.2}$$

where α_j is a taxon-specific intercept term, ϕ_j is the taxon-specific regression coefficient for treatment, and $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jP})'$ are the taxa-specific regression coefficients. The model cooperates with spike-and-slab priors on each of the regression coefficients. ϕ_j and θ_{jp}, with Gaussian slabs of mean 0 and variance r_j^2. They assumed Beta-Bernoulli priors for the latent inclusion indicators, $\varphi_j \sim \text{Beta-Bernoulli}\left(a_v, b_v\right)$ and $\zeta_{jp} \sim \text{Beta-Bernoulli}\left(a_t, b_t\right)$, respectively. In addition, $\alpha_j \sim \text{Normal}\left(0, \sigma_\alpha^2\right)$. The outcome model is a linear regression model that built the relationship between the outcome and the balance.

$$y_i = c_0 + c_1 t_i + \sum_{j=1}^{J-1} \beta_j B\left(\boldsymbol{\eta}_j, \boldsymbol{\psi}_i\right) + \sum_{p=1}^{P} \kappa_p z_{ip} + \epsilon_i \tag{28.3}$$

In the linear regression model, they assume the coefficients have spike-and-slab priors, that is

$$\beta_j \mid \xi_j, \sigma^2 \sim \xi_j N\left(0, h_\beta \sigma^2\right) + \left(1 - \xi_j\right) \delta_0\left(\beta_j\right), j = 1, \ldots, J-1, \tag{28.4}$$

$$\kappa_p \mid \nu_p, \sigma^2 \sim \nu_p N\left(0, h_\kappa \sigma^2\right) + \left(1 - \nu_p\right) \delta_0\left(\kappa_p\right), p = 1, \ldots, P, \tag{28.5}$$

They constructed the prior construction as $\xi_j \sim \text{Beta-Bernoulli}\left(a_j, b_j\right)$ and $\nu_p \sim \text{Beta-Bernoulli}\left(a_p, b_p\right)$, where (a_j, b_j) and (a_p, b_p) control the sparsity of the balances in the model. They assumed $c_0, c_1 \sim \text{Normal}\left(0, h_c \sigma^2\right)$ and $\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$, where $\sigma^2 \sim$Inverse-Gamma(a_0, b_0) for some $a_0 > 0$ and $b_0 > 0$.

28.2 Formula

Let y_i denote the observed continuous outcome of subject $i = 1, \ldots, n$ and $t_i \in \{0, 1\}$ the assigned treatment, with $t_i = 1$ if subject i received the treatment and $t_i = 0$ otherwise. Furthermore, let $z_i = (z_{i1}, \ldots, z_{iJ})'$ indicate a J-dimensional vector of taxa counts and $x_i = (x_{i1}, \ldots, x_{iP})'$ a P-dimensional vector of observed covariates. We first recast the joint model for compositional microbiome data of Koslovsky et al. (2020) into a framework for mediation analysis, where the relative abundances are treated as potential mediators, and then describe inference on direct and indirect effects following specification of the causal assumptions.

Bayesian Joint Model for Mediation Effect Selection

We adopt a joint model formulation that comprises a linear regression model for the phenotypic outcome and a Dirichlet-multinomial regression model for the compositional taxa. The two models

are linked via balances, calculated based on estimated relative abundances, that serve as the shared parameters.

Outcome Model: A multiple linear regression model is used to capture the direct effect of the treatment on the outcome, while adjusting for potential mediators and other covariates (including potential confounders of the outcome and mediators), as shown in Equation (1):

$$y_i = c_0 + c_1 t_i + \sum_{j=1}^{J-1} \beta_j B(\eta_j, \psi_i) + \sum_{p=1}^{P} \kappa_p x_{ip} + \epsilon_i, \tag{28.6}$$

where the balances $B(\eta_j, \psi_i)$ are a function of the relative abundances $\psi_i = (\psi_{i1}, \ldots, \psi_{iJ})'$, with $\sum_{j=1}^{J} \psi_{ij} = 1$. The relative abundances represent the proportion of each microbe in the microbial sample. Regression coefficients $\beta = (\beta_1, \ldots, \beta_{J-1})'$ represent the balances' effects, c_0 the intercept term, and c_1 the direct effect of treatment. Coefficients $\kappa = (\kappa_1, \ldots, \kappa_P)'$ capture the effects of the covariates x_i, and ϵ_i represents the error term. Spike-and-slab priors (Brown et al., 1998; George and McCulloch, 1997; Tadesse and Vannucci, 2021) are imposed on the coefficients β and κ, allowing us to investigate whether the balances and/or covariates are associated with the outcome, respectively. Specifically,

$$\beta_j | \xi_j, \sigma^2 \sim \xi_j N(0, h_\beta \sigma^2) + (1 - \xi_j)\delta_0(\beta_j), \quad j = 1, \ldots, J-1, \tag{28.7}$$

$$\kappa_p | \nu_p, \sigma^2 \sim \nu_p N(0, h_\kappa \sigma^2) + (1 - \nu_p)\delta_0(\kappa_p), \quad p = 1, \ldots, P, \tag{28.8}$$

where $\delta_0(\cdot)$ represents a Dirac delta function, or point mass, at zero. Here, the latent inclusion indicators ξ_j and ν_p take on values of 0 or 1, where $\xi_j = 1$ ($\nu_p = 1$) indicates that the corresponding balance (covariate) is included in the model, and 0 otherwise. We assume Bernoulli priors on the binary inclusion indicators, with Beta hyperpriors imposed on the inclusion probabilities. This allows the inclusion probabilities to be marginalized out for efficient sampling. We indicate this prior construction as $\xi_j \sim$ Beta-Bernoulli(a_j, b_j) and $\nu_p \sim$ Beta-Bernoulli(a_p, b_p), where a_j (a_p) and b_j (b_p) control the sparsity of the balances (covariates) in the model. To complete the outcome model's formulation, we assume $c_0, c_1 \sim$ Normal$(0, h_c\sigma^2)$ and $\epsilon_i \sim$ Normal$(0, \sigma^2)$, where $\sigma^2 \sim$ Inverse-Gamma(a_0, b_0) for some $a_0 > 0$ and $b_0 > 0$.

Dirichlet-Multinomial Model: The microbial taxa counts are treated as compositional and assumed to follow a multinomial distribution given the relative abundances ψ_i (i.e., $z_i \sim$ Multinomial$(z_i|\psi_i)$, where $z_i = \sum_{j=1}^{J} z_{ij}$). Conjugate priors for ψ_i can be specified as $\psi_i \sim$ Dirichlet(γ_i), where γ_i is a J-dimensional vector of concentration parameters. Note that the distributional assumptions for the taxa counts could take on various forms (Zhang et al., 2017). We chose a Dirichlet-multinomial (DM) model as it accommodates overdispersion and provides a computationally efficient Markov chain Monte Carlo (MCMC) routine that exploits data augmentation (Koslovsky et al., 2020). A log-linear regression framework can be used to relate the relative abundances with the treatment and covariates by introducing $\lambda_{ij} = \log(\gamma_{ij})$ and defining

$$\lambda_{ij} = \alpha_j + \phi_j t_i + \sum_{p=1}^{P} \theta_{jp} x_{ip}. \tag{28.9}$$

28.3 Pros and Cons

28.3.1 Pros

28.3.2 Cons

28.4 counterfactual assumption

28.5 Simulation

29 method

29.1 Formula

29.2 Pros and Cons

29.2.1 Pros

29.2.2 Cons

29.3 counterfactual assumption

29.4 Simulation

References