Testing Mediation Effect in High-dimensional Compositional Microbiome Data

Lei Liu

Division of Biostatistics
Washington University in St. Louis

co-authors: Haixiang Zhang, Jun Chen, Zhigang Li.

Outline

- Introduction
- Method
- Simulation
- Application
- Extension and Concluding Remarks

Microbiota

- Microbiota is a collection or community of microbes - microorganisms in a particular environment (including the body or a part of the body, e.g., gut, skin, vagina, or even urine)
- Some use "microbiome" to mean all the microbes in a community. Some use it as the genetic information of the microbiota
- Microbiota is associated with various diseases, e.g., obesity and diabetes (Everard and Cani 2013, Musso et al. 2010), Crohn's disease (Lewis et al. 2015), bacterial vaginosis (Srinivasan et al. 2012), and cancer (Garrett 2015, Schwabe and Jobin 2013).

Human Microbiome

- In 2007 the Human Microbiome Project was listed on the NIH Roadmap for Medical Research as one of the New Pathways to Discovery.
- More recently, microbiome has been found a key orchestrator of cancer therapy (Roy and Trinchieri 2017), especially cancer immunotherapy (Zitvogel et al. 2018): could the microbiome change the future of cancer treatment (Kruse 2018)?
- The microbiome is a key component of precision medicine (Petrosino 2018). For example, Zhu et al. (2018) demonstrated how precision editing of the gut microbiota may be used as a treatment for gastrointestinal inflammatory disease.

Microbial abundance

- Microbial abundance is usually measured in read counts. However, such quantities are not directly comparable across samples due to the uneven total sequence counts of samples.
- The read counts are normalized to relative abundances which sum to 1 for all microbes in a sample. For example, for the k-th sample, the compositional responses for p taxa are $(M_{k1}, M_{k2}, \ldots, M_{kp})^T$ where $M_{kj} \geq 0$ $(j = 1, 2, \ldots, p)$, and $\sum_{j=1}^{p} M_{kj} = 1$.

Challenges of Microbiome Data

- Compositional: each relative abundance is a non-negative value in [0, 1) which adds up to 1;
  - The p taxa $\mathbf{M} = (M_1, \cdots, M_p)'$ lies in a simplex (Aitchison, 1986)

$$\mathcal{S}^p = \left\{ \mathbf{x} = (x_1, \cdots, x_p)' : x_k > 0, k = 1, \cdots, p; \sum_{k=1}^{p} x_k = 1 \right\}.$$

  - Classical regression models in the real Euclidean space cannot be used to analyze the relative abundance directly (Aitchison 1999), e.g., for linear regression, any $p-1$ variables may contain the same information as all p variables.

- High dimensionality: the number of taxa p is high (487 in our application)
  - Variable selection is necessary
  - How to test a targeted taxon (say M_1) in the presence of high dimensional covariates (other taxa $M_2, \cdots, M_p$)?

Absolute Abundance vs. Relative Abundance

- The outcome: vote yes to a proposal in a state election (or GDP per capita)
- Covariates: numbers registered as Democrat, Republican, and Independent
- The absolute abundance does not matter much: California (with the largest population 39 million, thus large values of absolute abundances) does not have a much larger odds to vote yes than Wyoming (with the smallest population 586K, 1/67 of California)
- Rather, the composition (relative abundance) is more relevant

Compositional Feature

- Suppose there are 3 composition variables: $M_1 + M_2 + M_3 = 1$
- If we include all 3 variables as covariates:
  $E(Y) = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \beta_3 M_3$ is singular
- If we include only 2 variables
  - Model A: $E(Y) = \beta_0 + \beta_1 M_1 + \beta_2 M_2$
  - Because $M_2 = 1 - M_1 - M_3$, Model A can be rewritten as
    $E(Y) = (\beta_0 + \beta_2) + (\beta_1 - \beta_2) M_1 - \beta_2 M_3$
  - The coefficient of M_1 is now different: the parameter estimates and the inference statistics become useless, and an interpretation of these results is misleading.

Logratio Transformation for the Compositional Data

- Aitchison (1999) proposed the isometric logratio (*ilr*) transformation by transforming the compositional data from the simplex $\mathcal{S}^p$ to the Euclidean space $\mathbb{R}^{p-1}$ while preserving all metric (termed "isometric") properties.
- The *ilr* transformation on the compositional mediators $M_1, \cdots, M_p$ is

$$\tilde{M}_k = \sqrt{\frac{p-k}{p-k+1}} \ln \frac{M_k}{\sqrt[p-k]{\prod_{j=k+1}^{p} M_k}}, k = 1, \cdots, p-1. \qquad (1)$$

- ilr provides an orthonormal basis on the simplex, and uses the new coordinates in a standard linear regression model
- The only way to achieve a regression model without the need for constraints on the parameters, and with a meaningful interpretation of the unknown parameters.
- Note that two different ilr transformations, resulting in different orthonormal bases on the simplex, are orthogonal transformations of each other: invariance of the results of regression models on the choice of the orthonormal basis for the ilr transformation.

ilr of Targeted Variable $\tilde{M}_1$

- The transformed variable $\tilde{M}_1$ is a scaled sum of all logratios of original M_1 and $M_2, \cdots, M_p$, where the linear relationship is described as

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

- $\tilde{M}_1$ extracts all relative information of M_1 and captures the relative contribution of M_1 with respect to all the other parts (Hron et al., 2012)
- The interpretation of $\tilde{M}_1$ does not change if we were to permute $M_2, \cdots, M_p$
- In the previous example, does not matter which variable ($\tilde{M}_2$ or $\tilde{M}_3$) to include in the model if we are interested in $\tilde{M}_1$

Human Microbiome as Mediator

- We are interested in exploring the mediation mechanism of microbiome.
- Clinical question: how gut microbiota mediate the path from fiber intake to BMI (Wu et al. 2011)
- The fiber intake demonstrates a significant negative association with BMI, and the gut microbiota are significantly associated with both fiber intake and BMI (Zhang et al. 2018)
- Mediation analysis: fiber intake $\rightarrow$ gut microbiota $\rightarrow$ BMI.

Mediation Analysis

- Mediation analysis plays an important role in biomedical, behavioral, and psychosocial research studies.

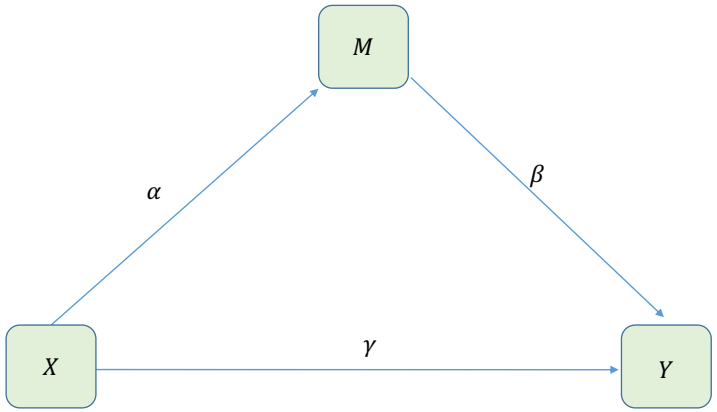

Figure 1. A scenario with a single mediator between exposure and outcome.

- Single mediator model:

$$\begin{aligned} Y &= c^* + \gamma^* X + \epsilon_1, \\ Y &= c + \gamma X + \beta M + \epsilon_2, \\ M &= c_1 + \alpha X + e_1, \end{aligned} \tag{2}$$

  - Y: the outcome
  - X: the exposure
  - M: the mediating variable or mediator
  - γ^*: represents the *total* effect of X on Y
  - γ: the *direct* effect of X on Y adjusted for the effect of the mediator M
  - α: relating the independent variable to the mediating variable
  - β: relating M to Y adjusted for the effect of X
  - *Indirect* effect: $\alpha\beta = \gamma^* - \gamma$

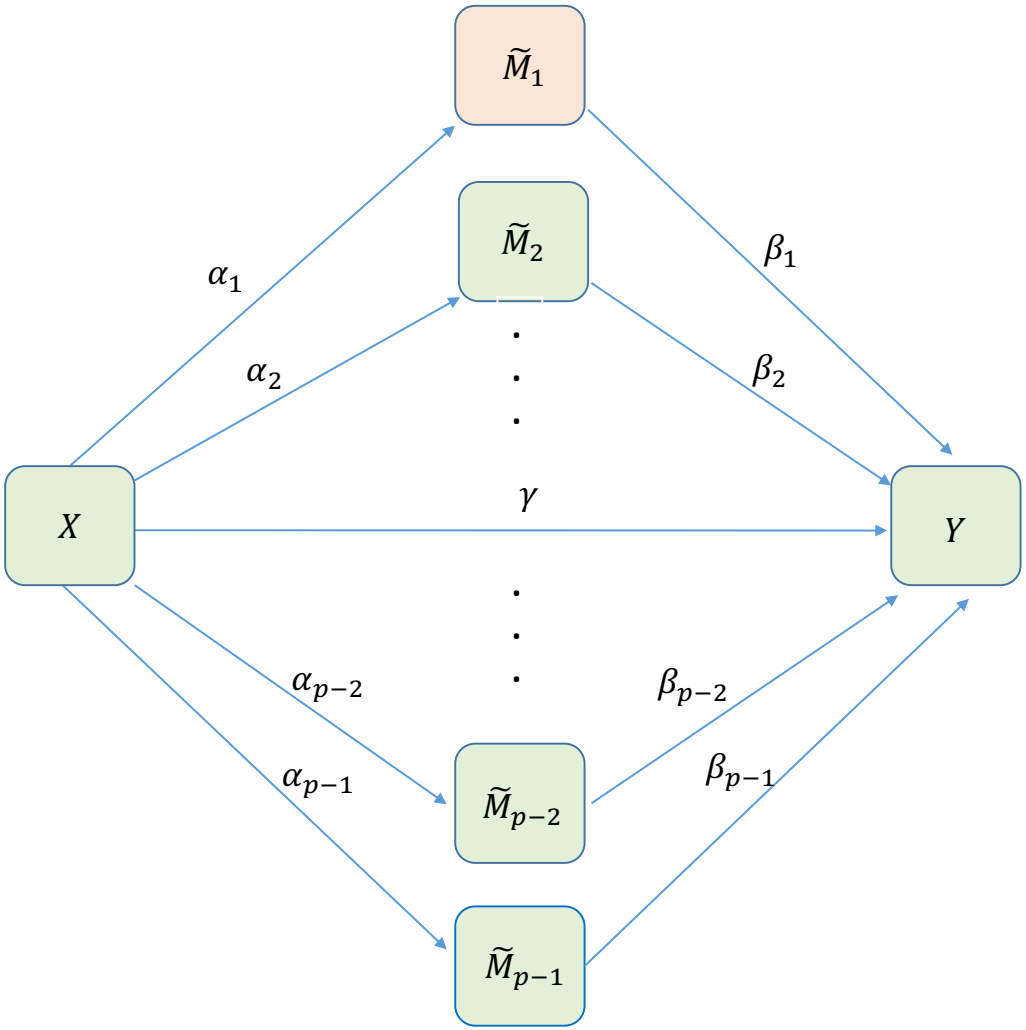

Figure 2. High-dimensional mediation model for the microbiome data, with the first coordinate $\tilde{M}_1$ being the targeted mediator.

Multiple (high-dimensional) Mediator Model in Microbiome

$$
\begin{aligned}
Y &= c^* + \gamma^* X + \epsilon_1, \\
Y &= c + \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{p-1} \tilde{M}_{p-1} + \epsilon_2, \\
\tilde{M}_k &= c_k + \alpha_k X + e_k, \quad k = 1, \cdots, p-1.
\end{aligned} \tag{3}
$$

- $\tilde{M}_k$: the kth (ilr) transformed mediating variable or mediator
- γ^*: represents the relation between X and Y
- γ: relating X to Y, adjusting for the effects of the mediators
- α_k: relating exposure to the kth mediating variable $\tilde{M}_k$
- β_k: relating $\tilde{M}_k$ to Y adjusting for the effect of X
- $\gamma^* = \gamma + \alpha_1\beta_1 + ... + \alpha_{p-1}\beta_{p-1}$

Naive Marginal Approach

- $Y = c + \gamma X + \beta_1 \tilde{M}_1 + \epsilon, j = 1, ..., p.$
- Not adjust for other mediators: Y depends on only one mediator $\tilde{M}_1$.
- Disadvantage of this method: Preacher and Hayes (2008)
  - In Figure 2, multiple mediators contribute to the outcome Y: imperative to adjust for other mediators in such analysis, especially given the potential correlations between different mediators.
  - Not feasible to predict Y using only one mediator (Zhang et al. 2016).

Three-step procedure

- **Step 1**: Conduct *ilr* transformation on the compositional mediators as in Equation (1).
- **Step 2**: Refit a linear regression model for $\tilde{M}_1$ as (3) in the Euclidean space.
- **Step 3**: Testing for the first *ilr* coordinate $\tilde{M}_1$,

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

For Other Mediators

- For mediator $\tilde{M}_k$ where $k \neq 1$, we can rearrange the order to make it the first coordinate, then run Steps 1-3.
- In a word, the first coordinate of the composition plays the role of *targeted* mediator.

Inference on the "Targeted" Mediation Effect

- Our aim is to estimate $\alpha_1\beta_1$ and construct the p-value for testing $H_0 : \alpha_1\beta_1 = 0$ vs. $H_1 : \alpha_1\beta_1 \neq 0$.
- For α_1, the ordinary least squares (OLS) estimator is denoted by $\hat{\alpha}_1$, and its corresponding variance estimate is $\hat{\sigma}^2_{\alpha_1}$.
- The OLS estimator of β_1 is not unique when the number of mediators p is larger than the sample size n.

De-biased Lasso estimator

- To solve this problem, we employ the de-biased Lasso technique (Zhang and Zhang 2014) to derive the estimator of β_1. Specifically, let

$$(\tilde{\gamma}, \tilde{\boldsymbol{\beta}}) = \arg\min_{\gamma, \boldsymbol{\beta}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \gamma X_i - \sum_{j=1}^{p-1} \tilde{M}_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \Big\}, \tag{4}$$

  where $\lambda > 0$ is the Lasso penalty parameter (Tibshirani, 1996).

De-biased Lasso estimator

- The de-biased Lasso estimator of β_1 is given by

$$\hat{\beta}_1 = \tilde{\beta}_1 + \frac{\sum_{i=1}^{n}(Y_i - \tilde{\gamma}X_i - \sum_{j=1}^{p-1}\tilde{M}_{ij}\tilde{\beta}_j)}{\sum_{i=1}^{n} Z_i \tilde{M}_{i1}}, \tag{5}$$

  - $\tilde{\gamma}$ and $\tilde{\boldsymbol{\beta}}$ are defined in (4).
  - $Z_i = \tilde{M}_{i1} - \hat{\eta}_1 X_i - \sum_{j=2}^{p-1}\hat{\eta}_j \tilde{M}_{ij}$, $k = 2, \cdots, p-1$.
  - $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \cdots, \hat{\eta}_{p-1})'$ is the Lasso solution from

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}\left(\tilde{M}_{i1} - \eta_1 X_i - \sum_{j=2}^{p-1}\eta_j \tilde{M}_{ij}\right)^2 + \lambda^* \sum_{j=1}^{p-1}|\eta_i| \right\}.$$

  - $\hat{\beta}_1$ is Lasso plus a one-step bias correction, and hence it is named "de-biased Lasso".

Asymptotic property (Zhang and Zhang 2014)

- It has been shown that $(\hat{\beta}_1 - \beta_{10})/\sigma_{\beta_1} \xrightarrow{\mathcal{D}} N(0,1)$
  - $\hat{\beta}_1$ is the de-biased Lasso estimator in (5)
  - The estimation of the standard error is given as

$$\hat{\sigma}_{\beta_1} = n^{-1/2} \frac{\hat{\sigma}_\epsilon \sqrt{\sum_{i=1}^n Z_i^2/n}}{|\sum_{i=1}^n Z_i \tilde{M}_{i1}/n|}, \tag{6}$$

  - $\hat{\sigma}_\epsilon^2 = \sum_{i=1}^n (Y_i - X_i\tilde{\gamma} - \sum_{j=1}^{p-1} \tilde{M}_{ij}\tilde{\beta}_j)^2/(n - \hat{s})$ (Reid et al. (2016))
  - $\hat{s}$ is the number of nonzero coefficients in the Lasso estimator $(\tilde{\gamma}, \tilde{\boldsymbol{\beta}})$

Joint significance test

- To test the targeted mediation effect $\alpha_1\beta_1$, we will adopt the *joint significance test*.
- The p-value is given by $P_{joint} = \max\{P_a, P_b\}$
  - $P_a = 2(1 - \Phi(|\hat{\alpha}_1|/\hat{\sigma}_{\alpha_1}))$; $P_b = 2(1 - \Phi(|\hat{\beta}_1|/\hat{\sigma}_{\beta_1}))$.
  - $\Phi(x)$ is the distribution function of $N(0, 1)$
  - $\hat{\alpha}_1$ and $\hat{\sigma}_{\alpha_1}$ are based on the OLS method;
  - $\hat{\beta}_1$ and $\hat{\sigma}_{\beta_1}$ are defined in (5) and (6), respectively.

Simulation

- Mediator $\tilde{M}$:
  - $X \sim N(0, 1.5)$
  - $c_k \sim U(1, 2)$
  - $\alpha = (\alpha_1, 0.15, 0.25, 0.35, 0.55, 0, \cdots, 0)'$ with $\alpha_1 = 0$, 0.10, 0.15, 0.25, 0.35, respectively;
  - $\mathbf{e} = (e_1, \cdots, e_p)'$ follows from $N(0, \Sigma)$. Here we consider three cases for the covariance matrix $\Sigma = (\Sigma_{ij})$
    - Case I: $\Sigma = I$;
    - Case II: $\Sigma_{jj'} = 0.75^{|j-j'|}$ for all $j, j' = 1, \cdots, p$;
    - Case III: $\Sigma_{13} = \Sigma_{31} = \Sigma_{35} = \Sigma_{53} = 0.5$; $\Sigma_{jj} = 1$ for all j and $\Sigma_{jj'} = 0$ for others.

- Outcome Y:
  - $c = 1$
  - $\gamma = 0.5$
  - $\epsilon_2 \sim N(0, 1)$
  - $\beta = (\beta_1, 0.25, 0.30, 0.35, 0.55, 0, \cdots, 0)'$ with $\beta_1 = 0$, 0.15, 0.25, 0.35.
- Other settings:
  - $p = 500$
  - Sample size $n = 100$ and 200, respectively.
  - 200 replicates.

Simulation Results

Table 1. BIAS and MSE (in parenthesis) for the $\alpha_1\beta_1$ with Case I†.

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	-0.0003	0.0025	-0.0001	0.0014
	(0.0076)	(0.0078)	(0.0042)	(0.0041)
(0.10, 0)	-0.0012	0.0092	0.0006	0.0058
	(0.0154)	(0.0159)	(0.0107)	(0.0091)
(0, 0.35)	-0.0004	0.0040	-0.0009	0.0005
	(0.0243)	(0.0215)	(0.0167)	(0.0168)
(0.15, 0.15)	-0.0021	0.0118	-0.0008	0.0060
	(0.0239)	(0.0258)	(0.0147)	(0.0148)
(0.25, 0.25)	0.0050	0.0227	0.0012	0.0104
	(0.0363)	(0.0388)	(0.0291)	(0.0231)
(0.35, 0.35)	-0.0002	0.0250	0.0013	0.0098
	(0.0487)	(0.0473)	(0.0356)	(0.0307)

† "Naive" is the marginal regression method.

Table 2. BIAS and MSE (in parenthesis) for $\alpha_1\beta_1$ with Case II†.

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	0.0020	0.0010	0.0004	0.0003
	(0.0506)	(0.0079)	(0.0346)	(0.0042)
(0.10, 0)	0.0657	0.0092	0.0705	0.0051
	(0.0503)	(0.0176)	(0.0349)	(0.0113)
(0, 0.35)	-0.0063	0.0012	-0.0028	0.0025
	(0.0717)	(0.0258)	(0.0448)	(0.0145)
(0.15, 0.15)	0.1040	0.0181	0.0990	0.0079
	(0.0598)	(0.0299)	(0.0419)	(0.0200)
(0.25, 0.25)	0.1673	0.0320	0.1724	0.0199
	(0.0655)	(0.0460)	(0.0510)	(0.0320)
(0.35, 0.35)	0.2422	0.0529	0.2349	0.0312
	(0.0824)	(0.0632)	(0.0628)	(0.0440)

† "Naive" is the marginal regression method.

Table 3. BIAS and MSE (in parenthesis) for $\alpha_1\beta_1$ with Case III†.

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	0.0010	0.0011	0.0007	0.0003
	(0.0129)	(0.0071)	(0.0085)	(0.0040)
(0.10, 0)	0.0143	0.0100	0.0152	0.0065
	(0.0192)	(0.0139)	(0.0123)	(0.0081)
(0, 0.35)	0.0036	0.0035	-0.0017	-0.0007
	(0.0325)	(0.0230)	(0.0263)	(0.0127)
(0.15, 0.15)	0.0218	0.0100	0.0249	0.0063
	(0.0303)	(0.0231)	(0.0215)	(0.0164)
(0.25, 0.25)	0.0365	0.0231	0.0403	0.0122
	(0.0464)	(0.0365)	(0.0295)	(0.0266)
(0.35, 0.35)	0.0520	0.0406	0.0495	0.0126
	(0.0637)	(0.0583)	(0.0398)	(0.0373)

† "Naive" is the marginal regression method.

Table 4.

Size and power at significance level 0.05 with Case I†.

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	0	0.005	0	0
(0.10, 0)	0.020	0.045	0.030	0.065
(0, 0.35)	0.030	0.025	0.055	0.055
(0.15, 0.15)	0.130	0.355	0.320	0.635
(0.25, 0.25)	0.510	0.860	0.790	0.970
(0.35, 0.35)	0.780	0.985	0.980	1

† "Naive" is the marginal regression method.

Table 5.

Size and power at significance level 0.05 with Case II†.

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	0.080	0	0.070	0
(0.10, 0)	0.270	0.015	0.575	0.055
(0, 0.35)	0.055	0.045	0.020	0.030
(0.15, 0.15)	0.595	0.285	0.845	0.410
(0.25, 0.25)	0.970	0.760	1	0.905
(0.35, 0.35)	1	0.940	1	0.995

† "Naive" is the marginal regression method.

Table 6.

Size and power at significance level 0.05 with Case III† .

	$n = 100$		$n = 200$	
(α_1, β_1)	Naive	Proposed	Naive	Proposed
(0, 0)	0.005	0	0.015	0
(0.1, 0)	0.075	0.035	0.205	0.055
(0, 0.35)	0.040	0.050	0.070	0.040
(0.15, 0.15)	0.355	0.295	0.775	0.600
(0.25, 0.25)	0.805	0.825	0.995	0.980
(0.35, 0.35)	0.955	0.985	1	0.995

† "Naive" is the marginal regression method.

Summary of Simulation Studies

- Our method is unbiased in all cases, while the Naive method is unbiased only in Case I with independent mediators
- From Tables 4-6, the Naive method yields inflated sizes when the mediators are correlated, which will result in too many false discoveries
- For our method, the sizes are conservative and consistent with the conclusion of the single mediator model with $\alpha_1 = \beta_1 = 0$ (MacKinnon et al. 2002).
- For ($\alpha_1 = 0$, $\beta_1 \neq 0$) or ($\alpha_1 \neq 0$, $\beta_1 = 0$), the sizes from our method are close to 0.05.

Application to Gut Microbiome data

- We apply our test procedure to a human gut microbiome data set, which includes 98 healthy subjects who were not on antibiotics for 3 months prior to data collection (Wu et al. 2011)
- We consider the fiber intake assessed by percent calories from dietary fiber (square-root transformed as in Zhang et al. 2018) as the exposure. Body mass index (BMI) was measured as the outcome.
- The fiber intake demonstrates a significant negative association with BMI, and the gut microbiota are significantly associated with both fiber intake and BMI (Zhang et al. 2018)
- Question of interest: fiber intake $\rightarrow$ gut microbiota $\rightarrow$ BMI.

- In between exposure and outcome, subjects' stool samples were collected and the DNA samples were analyzed by Roche 454 pyrosequencing of 16S rDNA gene segments. We thus have the abundance (count) of each taxon in the microbiome.
- Similar to Bokulich et al. (2013) and Yun et al. (2017), we removed a taxon if its total number in all samples is less than 0.04% of the grand total of all taxa in all samples, resulting in 487 taxa for analysis ($p > n$)
- Since the number of sequencing reads varied greatly across samples, these count data were transformed into compositions after zero counts were replaced by the maximum rounding error 0.5 (Lin et al. 2014; Cao et al. 2018). Thus, the potential mediators ($\mathbf{M}$) are compositional abundances of 487 taxa.
- To remove the compositional effects, we calculated the isometric logratio transformed $\tilde{\mathbf{M}}$ as in (1). For analysis, X and $\tilde{\mathbf{M}}$ are further standardized with mean 0 and variance 1.

Table 7.
Estimates and p-values of potential mediating taxa (Unadjusted p-value < 0.05)†.

ID	Phylum	Class	Order	Family	Genus	$\tilde{\alpha}$ (P_a)	$\tilde{\beta}$ (P_b)	P_{joint}
9441	F	C	C*	L	Other	−0.2002	1.2976	0.0453
						(0.0453)	(0.0321)	
98	F	C	C*	L	L*	0.3645	−1.5323	0.0304
						(0.0001)	(0.0304)	
14477	F	C	C*	V	Other	−0.2320	1.9022	0.0195
						(0.0195)	(0.0009)	
16444	F	C	C*	L	LIS	−0.2168	1.3478	0.0319
						(0.0296)	(0.0319)	

† $P_{joint} = \max\{P_a, P_b\}$; "F" denotes Firmicutes; "C" denotes Clostridia; "C*" denotes Clostridiales; "L" denotes Lachnospiraceae; "L*" denotes Lachnospira; "V" denotes Veillonellaceae; "LIS" denotes Lachnospiraceae Incertae Sedis.

Interpretation of Results

- We conduct mediation tests on individual taxon abundance by the proposed approach, where four taxa are significant with p-values smaller than 0.05 in Table 7.
- Specifically, the Lachnospira Genus has been proved to play an important role in the colonic fermentation of dietary fibers (Zhang et al. 2009).
- To adjust for multiple testings, we apply the FDR control. None of the taxa is significant under the FDR control, which is in line with the conclusion of Zhang et al. (2018).
- Although none of the associations survive multiple testing correction, the identified nominally significant taxa, coupled with strong biological evidence, justify a future large sample study.

Future Directions

- Phylogenetic tree structure.
  - Phylogenetic tree is a branching diagram or "tree" showing the evolutionary relationships among various biological species or other entities - their phylogeny - based upon similarities and differences in their physical or genetic characteristics.
  - Phylum, Class, Order, Family, Genus, Species (PC OF GS)
  - Taxa could co-exist or co-exclude: complicated covariance structure
- Zero inflation: two part model (Chen and Lee 2016, Chai et al. 2018).
- In addition to the structural equation modeling approach, the counterfactual approach of mediation analysis originated from causal inference should be considered

Major References

- Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, **26**, 35-67.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.
- Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, **76**, 217-242.
- Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L. and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, **32**, 3150-3154.
- Zhang, J., Wei, Z. and Chen, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*. In press.

Acknowledgements

- This work was supported by the Washington University Institute of Clinical and Translational Sciences grant UL1TR000448 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH)

Thank you!