Compositional Mediation Analysis for Microbiome Studies

Michael B. Sohn and Hongzhe Li

Department of Biostatistics and Epidemiology, University of Pennsylvania

Abstract

Motivated by advances in the causal inference in mediation analysis and problems arising in the analysis of metagenomic data, we consider the effect of treatment on an outcome transmitted through compositional mediators. Compositional and high dimensional features of such mediators make the standard mediation analysis not directly applicable. A sparse mediation model for high-dimensional compositional data is proposed in this paper utilizing the algebraic structure of a composition under the simplex space and a constraint linear regression model to achieve sub-compositional coherence. Under this model, we develop estimation method for estimating direct and microbial mediation effects of a randomly assigned treatment on an outcome. Tests for the total mediation effect of all bacterial taxa and individual mediation effects are also proposed. We conduct extensive simulation studies to assess the performance of the proposed method and apply the method to a real metagenomic dataset to investigate the effect of fat intake on body mass index (BMI) mediated through the gut microbiome composition.

1 Introduction

It has been shown that fat intake is associated with BMI (Bray and Popkin, 1998) and obesity is associated with the gut microbiome (Ley *et al.*, 2006; Turnbaugh *et al.*, 2006). From this information, a natural question to ask is whether fat intake has its effects on BMI mediated through the perturbation of the gut microbiome. The study of this kind of relationship is known as "mediation analysis". Mediation analysis is a statistical method of studying the effect of treatment or exposure on an outcome transmitted through intermediate variables, referred to as "mediators" or "intervening variables". It has been widely applied in various disciplines such as sociology, psychology, and epidemiology and become increasingly popular as the causal inference advances (Rubin, 2005; Imai *et al.*, 2010), which clarifies the assumptions needed for a causal interpretation. Until recently most of the mediation analysis has been restricted to a single mediator as depicted in Figure 1, and the effect of a mediator on an outcome is often formulated and implemented within the framework of linear structural equation models (LSEMs).

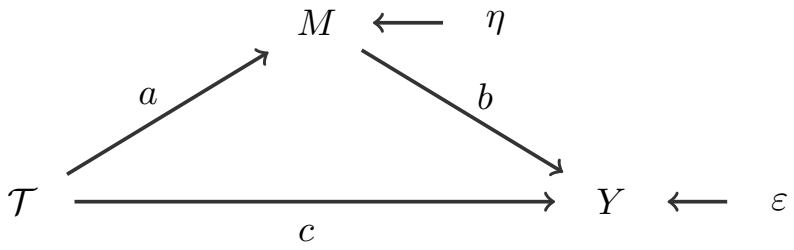

Figure 1: Path diagram for a single-mediator model: a, b and c are path coefficients and η and ε are random errors for a mediator M and an outcome Y, respectively.

For instance, LSEM for the path diagram in Figure 1 can be formulated as

$$M = a_0 + a\mathcal{T} + \eta,$$
$$Y = c_0 + c\mathcal{T} + bM + \varepsilon,$$

where $\mathcal{T}, M$, and Y are a treatment variable, a mediator variable, and an outcome variable; a, b, and c are path coefficients; η and ε are random errors for M and Y, respectively. Under this model, the effect of the treatment $\mathcal{T}$ on the outcome Y transmitted through the mediator M, often called *indirect effect*, is generally defined by the product of path coefficients a and b and the effect of $\mathcal{T}$ on Y not transmitted through M, *direct effect*, is defined by the path coefficient c. It is easy to see that under LSEM the total effect of $\mathcal{T}$ on Y is the sum of direct and indirect effects.

In recent years, numerous studies have attempted to extend the applicability of mediation analysis:

incorporating nonlinearity and interactions (Pearl, 2001; Imai *et al.*, 2010; VanderWeele and Vansteelandt, 2010) and including multiple mediators (Preacher and Hayes, 2008; VanderWeele and Vansteelandt, 2014). There have also been a few studies to develop mediation analyses for high-dimensional mediators. Chén *et al.* (2015) proposed a method to estimate the path coefficients by finding linear combinations of mediators that maximize the likelihood of LSEM, which is similar to the principal components. Huang and Pan (2016) also introduced a transformation model using the principal components and included interaction terms in their model. Zhao and Luo (2016) proposed a sparse mediation model using a regularized LSEM approach.

This paper considers the problem of mediation analysis when mediators are compositional data, which are often observed in the microbiome and metagenomic studies. Compositional data refer to proportions or percentages of a whole and arise frequently in a wide range of disciplines such as mineral components of a rock in geology, vote shares of an election in psephology, and microbial components of a natural environment in biological science. In microbiome studies, to account for different sizes of sequencing libraries for either 16S rRNA sequencing or shotgun metagenomic sequencing, the sequencing reads count data are often normalized into proportions. That the components of a composition sum to unity makes a radical difference between the sample space for compositional data, which is known as $(k-1)$-dimensional simplex ∇^{k-1} for k components, and the real Euclidean space $\mathbb{R}^k$ associated with unconstrained data, thus making the models for unconstrained data inappropriate for compositional data. To deal with this special nature of compositional data, Aitchison (1982) introduced an axiomatic approach with a variety of operations under logratio transformation, which provides a one-to-one mapping between ∇^{k-1} and $\mathbb{R}^k$, and various researchers including himself have formalized and extended his approach. Aitchison and Bacon-Shone (1984) proposed linear and quadratic log contrast modes for compositional covariates and Billheimer *et al.* (2001) formulated the algebraic structure of a composition under the simplex space. Lin *et al.* (2014) and Shi *et al.* (2016) developed and generalized high-dimensional log contrast modes for compositional data regression, respectively.

This paper develops a framework for mediation analysis when mediators are high dimensional compositional data, as often observed in microbiome and metagenomic studies. Our mediation framework includes two components. First, the compositional algebra of Billheimer *et al.* (2001) is used to quantify the effect of a randomly assigned treatment on overall microbiome composition. Specifically, we propose a method that minimizes a compositional norm to estimate the treatment effect on microbiome composition. Second, the association between an outcome, treatment and microbial composition is modeled using the compositional data regression in a high dimensional setting (Lin *et al.*, 2014; Shi *et al.*, 2016). Under this modeling framework, both direct and microbial mediation effects of treatment on an outcome can be quantified and estimated. Additionally, tests for a total microbiome mediation effect and each individual taxon mediation ef-

fect are developed using a bootstrap procedure or extending the Sobel test (Sobel, 1982) for high-dimensional mediators.

The rest of this paper is organized as follows. Section 2 introduces the compositional mediation model for a continuous outcome and discusses model assumptions and identifiability. Section 3 proposes methods of estimating parameters and their covariance matrices. We also discuss null hypotheses for a total mediation effect and component-wise mediation effects. Section 4 compares the performance of our method with two other methods that can be applied to compositional data in extensive simulation studies. Section 5 presents an application of the proposed methods to gut microbiome data. Finally, Section 6 presents a brief discussion of the model and its potential extensions.

2 Compositional Mediation Model and Causal Interpretation

2.1 Compositional Mediation Model

Suppose that we have n independent identically distributed (iid) samples, each consisting of an outcome y_i, a treatment τ_i, and k-dimensional compositional covariates M_i, where $i = 1, \ldots, n$. Here M_i lies in the $(k-1)$ dimensional simplex ∇^{k-1} for all i, that is, $M_i = \{(M_{i1}, \ldots, M_{ik}) : M_{ij} > 0,\ j = 1, \ldots, k,\ \sum_{j=1}^{k} M_{ij} = 1\}$. A model describing the effect of τ on y mediated through M is schematically depicted in Figure 2.

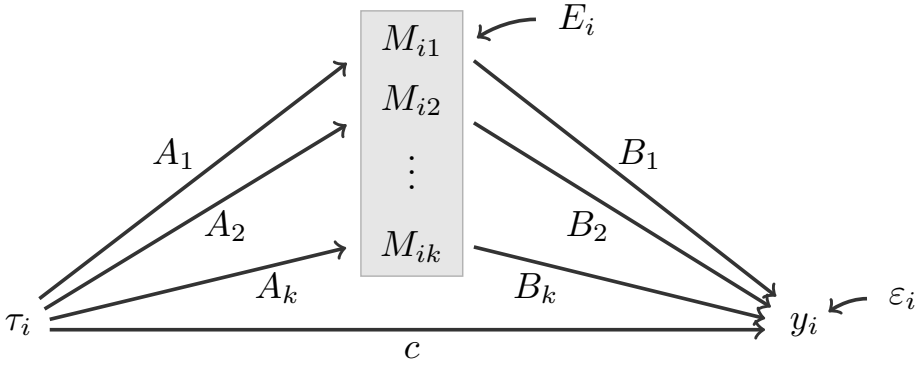

Figure 2: A compositional mediation model: A_j, B_j and c are path coefficients, $j = 1, \ldots, k$. E_i and ε_i are random errors for a k-component compositional mediator M_i and an outcome y_i, respectively. $\sum_{j=1}^{k} M_{ij} = 1$.

Since M_i lies in the $(k-1)$ dimensional simplex ∇^{k-1}, multivariate regression is inappropriate to estimate the effect of τ on M, which assumes M_i lies in the k-dimensional Euclidean space $\mathbb{R}^k$. For the same reason, multiple regression is not appropriate to estimate the effect of τ and M on y. Before we introduce the compositional mediation model, we define two compositional operators. For two compositions $\boldsymbol{\eta}, \boldsymbol{\zeta} \in \nabla^{k-1}$,

a perturbation operator is defined by

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

and a power transformation for a composition $\boldsymbol{\eta}$ by a scalar α by

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

To properly account for the nature of compositional data, we propose the following compositional mediation model:

$$M_i = \left(M_0 \oplus A^{\tau_i^*} \right) \oplus E_i \tag{1}$$

$$y_i = c_0 + \tau_i^* c + (\log M_i)^T B + \varepsilon_i, \quad \text{subject to } B^T \mathbf{1}_k = 0, \tag{2}$$

where $\tau_i^* = \tau_i - \bar{\tau}$, where $\bar{\tau}$ is the mean of τ; M_0 is a vector of compositions at the baseline level (i.e., when $\tau_i = \bar{\tau}$); E_i is a vector of the random composition perturbation errors; $\varepsilon_i \sim N(0, \sigma^2)$ is the random error of the outcome; $\mathbf{1}_k$ is a k column vector of ones; A, B, c_0, and c are parameters needed to be estimated. The superscript $\top$ denotes the transpose operator. Model (1) models how a treatment τ_i^* perturbs a compositional mediator M_i from the baseline composition M_0, which is measured by the compositional effect A. Model (2) links a compositional mediator M_i and a treatment τ_i^* to an outcome y_i. In order to take the compositional nature of M_i into account, a linear constraint $B^T \mathbf{1}_k = 0$ is imposed so that the model is subcompositional invariant (Lin *et al.*, 2014; Shi *et al.*, 2016).

In this compositional mediation model, the total effect of τ on y can be decomposed into a direct effect c and a total indirect effect $[\log(kA)]^T B$ or $(\log A)^T B$, which is the sum of component-wise indirect effects. Note that under the model (1), A is interpreted as the expected change of M from the identity element $\mathcal{J}_{k-1}$ in ∇^{k-1}, where $\mathcal{J}_{k-1} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$. Thus, A is divided by $\mathcal{J}_{k-1}$, or multiplied by k, so that $\log(kA_j)$ represents the expected change of $\log(M_j)$ from 0 by one unit change of τ for the j^{th} compositional mediator.

2.2 Model assumptions for causal interpretation

The identification of causal direct and indirect effects requires several assumptions. Let X denote a vector of covariates that are non-descendants of a treatment $\mathcal{T}$ and a compositional mediator M. Then, after

controlling for X, the assumptions under the potential outcome framework can be given by

$$\{Y(\tau', m), M_j(\tau)\} \perp\!\!\!\perp \mathcal{T}|X, \tag{3}$$

$$Y(\tau', m) \perp\!\!\!\perp M_j(\tau)|(\mathcal{T}, X), \tag{4}$$

for all $j = 1, \ldots, k$. These assumptions are an extension of the sequential ignorability assumptions for the single mediator model (Imai *et al.*, 2010). The first assumption states that, given the observed covariates, there is no unmeasured confounder associated with the treatment and the outcome and with the treatment and each component of the compositional mediator. The second assumption states that there is no unmeasured confounder associated with each component of the compositional mediator and the outcome given the treatment and the observed covariates. These assumptions seem to be similar to a multivariate extension of those for the single mediator model, which in general makes the assumptions much stronger (VanderWeele and Vansteelandt, 2014). However, the assumptions (3) - (4) are not as strong as those for a multiple-mediators model due to correlations among compositional mediators.

Theorem 1. (Model assumption for causal interpretation) *If a compositional mediator M_j satisfies the assumptions (3) - (4), then other compositional mediators M_i for all $i \neq j$ satisfy the assumptions.*

The proof of Theorem 1 is given in Appendix A. No interactions between the treatment and the mediators and among the mediators are also assumed, which is a standard assumption in causal mediation inference for LSEM. Under these assumptions, we have the following theorem to show that models (1) and (2) lead to quantification of direct and indirect mediation effects.

Theorem 2. (Identification for the compositional mediation model) *Under the assumptions (3) - (4) and no interaction assumption, a direct effect $\delta(\tau^*)$ and a total indirect effect $\zeta(\tau^*)$ for the compositional mediation model (1) - (2) are identifiable and given by*

$$\delta(\tau^*) = c, \tag{5}$$

$$\zeta(\tau^*) = (\log A)^T B. \tag{6}$$

The proof of Theorem 2 is given in Appendix B.

3 Parameter Estimation, Variance Estimation, and Tests of Mediation Effects

3.1 Estimation of composition parameters and covariance matrix

With the compositional operators, Billheimer *et al.* (2001) show that ∇^{k-1} constitutes a complete inner product space, allowing the definition of a norm for a composition vector $\boldsymbol{\eta}$, which is given by

$$\|\boldsymbol{\eta}\| = \left(\boldsymbol{\eta}^T\boldsymbol{\eta}\right)^{1/2} = \left(\boldsymbol{\theta}^T\mathcal{N}^{-1}\boldsymbol{\theta}\right)^{1/2},$$

where $\boldsymbol{\theta} = \mathrm{alt}(\boldsymbol{\eta})$ and

$$\mathcal{N}^{-1} = I_{k-1} - \frac{1}{k}\mathbf{1}_{k-1}\mathbf{1}_{k-1}^T,$$

where I_{k-1} is a $(k-1)$-dimensional identity matrix.

To estimate the parameters in Model (1), we propose minimizing the norm of the difference between observed and predicted compositions, that is,

$$\begin{aligned}\hat{A} &= \underset{A, M_0 \in \nabla^{k-1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| M_i \ominus M_0 \oplus A^{\tau_i^*} \right\|^2 \\ &= \underset{A, M_0 \in \nabla^{k-1}}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{k-1} \left\{ (k-1)\left[\log\left(\frac{M_{ij}M_{0k}A_k^{\tau_i^*}}{M_{ik}M_{0j}A_j^{\tau_i^*}}\right)\right]^2 \right. \\ &\qquad \left. - \log\left(\frac{M_{ij}M_{0k}A_k^{\tau_i^*}}{M_{ik}M_{0j}A_j^{\tau_i^*}}\right) \sum_{\ell \neq j}^{k-1} \log\left(\frac{M_{i\ell}M_{0k}A_k^{\tau_i^*}}{M_{ik}M_{0\ell}A_\ell^{\tau_i^*}}\right) \right\},\end{aligned} \tag{7}$$

where an inverse operator $\ominus$ is defined by

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

The objective function Eq. (7) is not convex. However, since the square of a logarithmic function in the form of $\log(ax^b)$, where $a > 0$ and $b \neq 0$ are constants and $x \in \mathbb{R}^+$, has only one stationary point (i.e., a point at which the derivative of a function vanishes) and the minimum occurs at this point, its optimal solution

(i.e., a solution of an M-estimator) can be obtained by solving the system of linear equations,

$$\begin{bmatrix} (k-1)c_{\tau^*} & -c_{\tau^*} & \dots & -c_{\tau^*} \\ -c_{\tau^*} & (k-1)c_{\tau^*} & \dots & -c_{\tau^*} \\ \vdots & \vdots & \ddots & \vdots \\ -c_{\tau^*} & -c_{\tau^*} & \dots & (k-1)c_{\tau^*} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{k-1} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{k-1} \end{bmatrix}$$

with a constraint $A^T\mathbf{1}_k = 1$, where $c_{\tau^*} = \sum_{i=1}^n \tau_i^{*2}$, $f_j = \log(A_j/A_k)$, and $m_j = k\sum_{i=1}^n \tau_i^* \log M_{ij} - \sum_{i=1}^n \sum_{\ell=1}^k \tau_i^* \log M_{i\ell}$. The matrix in the above equation is always invertible by Sherman-Morrison formula (Sherman and Morrison, 1950) since it can be expressed as the sum of a diagonal matrix of kc_{τ^*} and a constant matrix of $-c_{\tau^*}$. Note that in Model (1) and (2), for brevity, we assume no confounding variable X and exclude it. However, including X in the models does not add much complexity. For instance, if n_x confounding variables are included, the system of $n_x(k-1)$ linear equations needs to be solved for the model (1); see Appendix C for details.

In Model (1), we do not specify the distribution of E_i. Therefore, we use a bootstrap distribution of $\hat{A}$ to estimate its covariance matrix $\hat{\Sigma}_A$, specifically using the percentile method (Machado and Parente, 2005). We define a matrix H as

$$H = \left[\begin{array}{c|c} I_k & 0 \\ \hline C & D_{-2} \end{array}\right],$$

where I_k is a $k \times k$ identity matrix; 0 is a $k \times k(k-1)/2$ matrix of zeros; C is a $k(k-1)/2 \times k$ binary matrix whose rows are all possible combinations of 2 ones and $k-2$ zeros; D_{-2} is a $k(k-1)/2 \times k(k-1)/2$ scalar matrix with -2. Then, $\hat{\Sigma}_A$ can be obtained by solving the system of linear equations,

$$H\boldsymbol{u} = \boldsymbol{v},$$

where $\boldsymbol{u} = (\hat{\sigma}_{11}, \dots, \hat{\sigma}_{kk}, \hat{\sigma}_{12}, \dots, \hat{\sigma}_{(k-1)k})^T$ is a vector of all the unique elements in a covariance matrix of $\hat{A}$ and $\boldsymbol{v} = (s^2_{h_1}, \dots, s^2_{h_{k(k+1)/2}})^T$ is a vector of bootstrap estimators of $\hat{\sigma}_h$. For instance, from the first, second, and $(k+1)$th rows, we can obtain $\hat{\sigma}_{11} = s^2_{h_1}$, $\hat{\sigma}_{22} = s^2_{h_2}$, and $\hat{\sigma}_{11} + \hat{\sigma}_{22} - 2\hat{\sigma}_{12} = s^2_{h_{k+1}}$ or $\hat{\sigma}_{12} = (s^2_{h_1} + s^2_{h_2} - s^2_{h_{k+1}})/2$. For the estimators of $\hat{\sigma}_H$, we use functionals censored at the top and bottom 5% of the bootstrap distribution,

$$s_{h_i} = \frac{1}{0.847 n_b} \sum_{j=s}^{\ell} \Phi^{-1}(j/n_b) q^*_{h_i}(j/n_b) + 0.103 \left[q^*_{h_i}(\ell/n_b) - q^*_{h_i}(s/n_b)\right]$$

with $s = \lfloor 0.05 n_b \rfloor$ and $\ell = B - \lfloor 0.05 n_b \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x; n_b is a bootstrap sample size;

$\Phi^{-1}(\alpha)$ is the α-percentile of the standard normal distribution; $q^*_{h_i}(\alpha)$ is the α-percentile of the bootstrap distribution of $\left\{\sqrt{n}(\boldsymbol{h}_i^T \hat{A}_b^* - \boldsymbol{h}_i^T \hat{A}), b = 1, \ldots, n_b\right\}$. $\hat{A}_b^*$ is a vector of bootstrap estimations and $\boldsymbol{h}_i$ is a vector satisfying $\boldsymbol{h}_i^T \hat{\Sigma}_A \boldsymbol{h}_i = \hat{\sigma}_{h_i}$. For the $(k+1)$th row, as an example, $\boldsymbol{h}_{k+1} = (1, -1, 0, \ldots, 0)^T$.

3.2 Estimation of compositional regression parameters and covariance matrix

Aitchison and Bacon-Shone (1984) introduced linear log-contrast models for a mixture of k components,

$$y_i = Z_i^T \boldsymbol{\beta} + \varepsilon_i, \text{ subject to } \sum_{j=1}^{k} \beta_j = 0, \tag{8}$$

where $Z_i = \{(\log P_{i1}, \ldots, \log P_{ik}) : P_{ij} > 0,\ j = 1, \ldots, k,\ \sum_{j=1}^{k} P_{ij} = 1\}$. Note that for simplicity the intercept is excluded in the model, which can be eliminated by centering all the variables in the model. An ℓ_1 penalization method for high-dimensional linear log-contrast models was proposed by Lin *et al.* (2014), and extended by Shi *et al.* (2016) to accommodate subcompositional log-contrasts and estimate de-biased ℓ_1 regularized estimates and their covariance matrix.

Model (2) can be analyzed by a method for the linear log-contrast model. Specifically, define

$$\widetilde{Z}_i = Z_i^T \left(I_{k+1} - \tilde{\mathbf{1}}_{k+1} \tilde{\mathbf{1}}_{k+1}^T / k\right), \tag{9}$$

where $Z_i = (\log M_{i1}, \ldots, \log M_{ik}, \tau_i^*)^T$ and $\tilde{\mathbf{1}}_{k+1} = (\mathbf{1}_k, 0)^T$. Then, Model (2) can be expressed as

$$y_i = \widetilde{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \text{ subject to } \sum_{j=1}^{k} \beta_j = 0, \tag{10}$$

where $\boldsymbol{\beta} = (B_1, \ldots, B_k, c)^T$, that is, we solve the penalized constrained convex optimization problem,

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2n} \left\| y - \widetilde{Z}\boldsymbol{\beta} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1\right), \text{ subject to } \sum_{j=1}^{k} \beta_j = 0. \tag{11}$$

As proposed by Shi *et al.* (2016), we then apply the de-biasing procedure to the solution of Equation (11) to compensate the bias introduced by the ℓ_1 penalty and obtain de-biased estimates and their covariance matrix.

3.3 Hypothesis test of mediation effect

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

$$H_0 : (\log A)^T B = 0, \tag{12}$$

and the null hypothesis of no component-wise mediation effects is given by

$$H_0 : (\log kA_j)B_j = 0, \ \forall j \in \{1, 2, \ldots, k\}. \tag{13}$$

Null Hypothesis (12) reflects a total mediation effect on an outcome that we could observe; however, it can disguise actual mediation effects, which can be captured by Null Hypothesis (13). If for instance, the magnitude of mediation effects of M_i and M_j are equal, but their direction is opposite, then a total mediation effect is zero. In other words, Null Hypothesis (12) will not be rejected while Null Hypothesis (13) will be. Therefore, both need to be tested to interpret mediation effects properly.

To test Null Hypothesis (12) and (13), we propose two approaches: an extension of the Sobel test (Sobel, 1982) for high-dimensional mediators and a bootstrap procedure. To test Null Hypothesis (12) with the former, the square root of the first order asymptotic variance of a total indirect effect, which is computed with estimated covariance matrices of $\log(k\hat{A})$ and $\hat{B}$ by the method described in (Bollen, 1987), is used as a standard error for the total indirect effect in the Z-test. The formulae for the first order asymptotic variance of a total indirect effect and component-wise indirect effects are given in Appendix D.

Note that our estimator for a total indirect effect $\zeta(\tau^*)$ is an unbiased estimator since our model is under the framework of LSEM as shown in the proof of Theorem 2. Also, the distribution of compositions M_i is not known, but the log-ratio of compositions (e.g., $\log M_{ij}/M_{ik}$) is well approximated by a normal distribution (Aitchison, 1986), that is, the distribution of $\log(\hat{A}_j/\hat{A}_k)$ is well approximated by a normal distribution. Therefore, $\zeta(\tau^*)$ can also be approximated by a normal distribution assuming the product of two normal variables approximately follows a normal distribution. In general, the product of two normal variables does not follow a normal distribution. However, a misspecified distribution of the product of two normal variables will just reduce the power to detect the indirect effects when the null hypothesis of no indirect effect is false, not affecting type I error rates (MacKinnon *et al.*, 2002; Shrout and Bolger, 2002).

A bootstrap approach can be used to avoid the assumptions of normality for the product of two normal variables (Shrout and Bolger, 2002; VanderWeele and Vansteelandt, 2014). To this end, we use a non-parametric bootstrap for $\log(k\hat{A})$ and a parametric bootstrap for $\hat{B}$ using the estimated multivariate normal distribution to approximate the sampling distribution of $\zeta(\tau^*)$. The p-value for $\zeta(\tau^*)$ is then approximated

by utilizing the fact that any bootstrap replicate $\zeta(\tau^*)_b - \zeta(\tau^*)$ should have a distribution close to that of $\zeta(\tau^*)$ when the null hypothesis is true, where $\zeta(\tau^*)_b$ denotes the estimated total indirect effect derived from the bootstrap samples (Efron and Tibshirani, 1993). In other words, the p value can be approximated by $2\sum_{b=1}^{n_b} \mathrm{I}(\zeta(\tau^*)_b - \zeta(\tau^*) > \zeta(\tau^*))/n_b$ when $\zeta(\tau^*) \geq 0$ and $2\sum_{b=1}^{n_b} \mathrm{I}(\zeta(\tau^*)_b - \zeta(\tau^*) < \zeta(\tau^*))/n_b$ when $\zeta(\tau^*) < 0$, where $\mathrm{I}(\cdot)$ is the indicator function and n_b is a number of bootstrap samples. Similarly, Null Hypothesis (13) can be tested. In the simulation, we observed no significant difference in performance between the two methods in terms of power and type I error, concurring with Huang and Pan (2016).

4 Simulation Studies

Mediation analysis for multiple or high dimensional mediators typically assumes independence between mediators to establish a causal interpretation so principal components of mediators are often used. This principal component regression (PCR) approach is also applicable for compositional mediators to estimate a total indirect effect (TIDE). Another possible approach for compositional mediators is utilizing an ℓ_1 regularization, which has a tendency of dropping correlated variables. We used these two approaches to evaluate the performance of our compositional mediation model (CMM). For the hypotheses tests, we used the extension of the Sobel test for fair comparison.

In data generation, a treatment variable τ was generated by a normal distribution, $\tau_i \sim N(0,1)$. A vector of compositions at the baseline level M_0 was randomly generated with the unit-sum constraint. The parameters A, B and c were selected such that a total indirect effect is approximately 0.00, 0.50, 0.75 or 1.00 and a direct effect is 1.00. For the random errors, we used $E_i \sim LN(\mathbf{0}_{k-1}, 2\mathcal{N})$ and $\varepsilon_i \sim N(0,2)$, where LN denotes a logistic normal distribution and $\mathcal{N} = \left[I_{k-1} + \mathbf{1}_{k-1}\mathbf{1}_{k-1}^T\right]$. Note that our model does not specify the distribution of E_i. A vector of compositions M_i and an outcome y_i were generated according to Model (1) and Model (2), respectively. All variables were centered so the intercept $c_0 = 0$.

We first estimated the power and type 1 error rate of testing the TIDE from 1500 simulations for each k mediators with a sample size $n = 100$ at various probability thresholds, where $k = 5, 49, 99$: 250 simulations with each of TIDE $= 1.00, 0.75, 0.50$ and 250 simulations with each of the following: no effect of τ on M (i.e., $A_j = 0, \forall j$), no effect of M on y (i.e., $B_j = 0, \forall j$) and *inconsistent* TIDEs (i.e. the sum of component-wise indirect effects is zero). For the PCR approach, only the first k_{pc} principal components that explain 90% of the total variance were included. As an ℓ_1 regularization approach, we used the two stage adaptive Lasso (TSAL), which uses the standard Lasso in the first stage to screen irrelevant variables and the adaptive Lasso in the second stage to select consistent variables (Bühlmann and van de Geer, 2011). As shown in Table 1, while all three methods roughly control the type 1 errors, CMM outperforms both the PCR and

TSAL approaches in power, especially when k is large.

Table 1: Power and type 1 error rate of testing TIDE for $k = 5, 49, 99$ with a sample size $n = 100$ at the significance level $\alpha = 0.001, 0.01, 0.05$. A total of 1500 simulations for each k were used: 750 simulations with non-zero TIDE and 750 simulations for zero TIDE. CMM: proposed compositional mediation model; PCR: principal component regression; TSAL: two-stage adaptive Lasso.

		Power			Type 1 error		
α		0.001	0.01	0.05	0.001	0.01	0.05
$k = 5$	CMM	0.452	0.623	0.752	0.001	0.011	0.035
	PCR	0.381	0.604	0.732	0.000	0.007	0.040
	TSAL	0.413	0.621	0.740	0.001	0.008	0.037
$k = 49$	CMM	0.476	0.675	0.820	0.001	0.015	0.043
	PCR	0.051	0.207	0.423	0.000	0.005	0.029
	TSAL	0.277	0.487	0.631	0.001	0.012	0.067
$k = 99$	CMM	0.397	0.645	0.791	0.001	0.007	0.040
	PCR	0.023	0.105	0.273	0.000	0.004	0.028
	TSAL	0.304	0.495	0.628	0.003	0.016	0.064

We then examined the estimated direct and indirect effects at various sample sizes to see the bias and variance of the three methods. Figure 3 shows the results for $k = 49$ based on 100 simulations. When the sample size is small, the estimates obtained by CMM are slightly biased. However, as the sample size increases, the estimates obtained by all three methods converge to the true values.

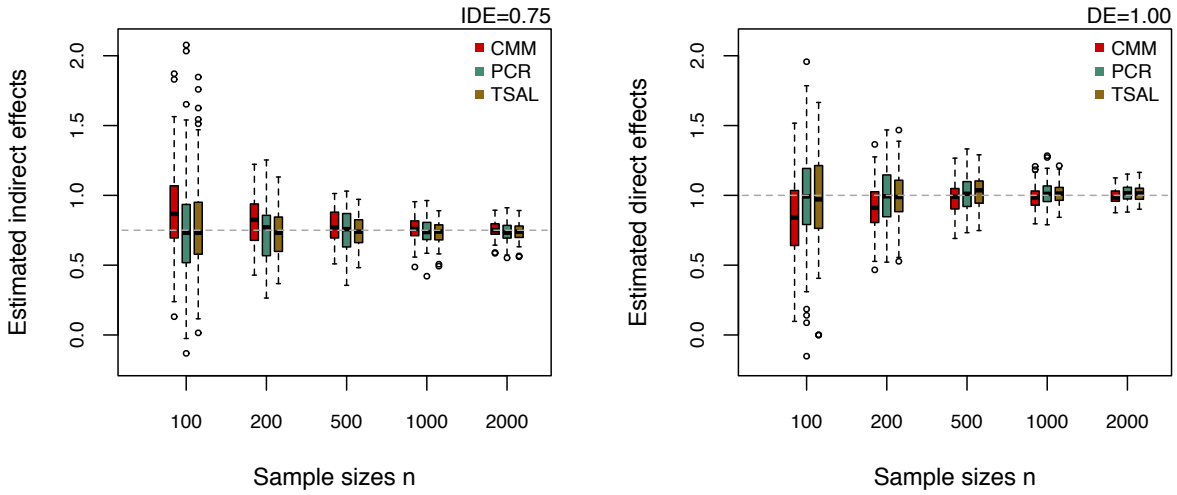

Figure 3: Estimated direct and indirect effects at various sample sizes for $k = 49$ with a direct effect of 1.00 and an indirect effect of 0.75. The results are based on 100 simulations. CMM: proposed compositional mediation model; PCR: principal component regression; TSAL: two-stage adaptive Lasso.

The PCR approach is not capable of testing component-wise IDEs. Therefore, we compared the performance of CMM on component-wise IDEs just with TSAL. In data generation, we selected the parameters A and B such that the first 7 IDEs are at a non-zero constant level with varying values of B and the remaining 43 IDEs are at zero. The level of IDEs was increased by 0.1 from 0.0, and the sample size of 100 and 200 were used. For multiple testing corrections, we used false discovery rate (FDR) using the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) at 0.05. As a comparison measure, the F_1 score, which is the harmonic mean of precision and recall, was used. Figure 4 shows results: better performance of CMM over TSAL.

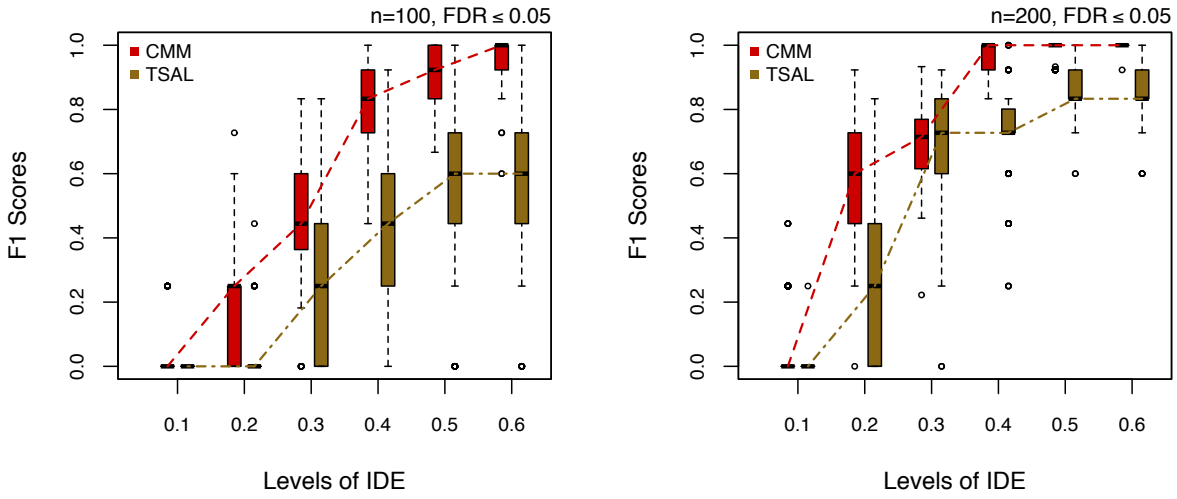

Figure 4: F_1 score versus levels of IDEs at FDR ≤ 0.05. P-values are adjusted by FDR using the Benjamini-Yekutieli procedure. CMM: proposed compositional mediation model; TSAL: two-stage adaptive Lasso.

5 Real Data Application

We applied CMM to the dataset reported in Wu *et al.* (2011), which is the source of motivation for this study. The dataset consists of 16S rRNA sequences from fecal samples of 98 healthy volunteers from the University of Pennsylvania. It also contains demographic and clinical information including fat intakes and BMI. We summarized OTUs at the genus level and then filtered out the genera that appear less than 10% of the samples, leaving 45 genera in 96 samples. Due to different total counts throughout the samples, the OTU counts assigned to these genera were transformed into proportions after replacing zero counts by the maximum rounding error 0.5, which is commonly used in compositional data analysis (Aitchison, 1986).

The gut microbiota can influence host adiposity through energy extraction from the diet, with variable

efficiency depending on community composition; furthermore, the microbiota can also affect host adiposity by influencing metabolism throughout the body. It is therefore highly likely that gut microbiome can potentially mediate the effect of diets such as fat on host adiposity and BMI. Since the 98 samples were roughly randomly sampled, it was reasonable to assume that fat was randomly sampled. CMM was applied to the dataset with BMI as the outcome, fat intake as the treatment, and the 45 genera as the compositional mediators. The estimated DE and TIDE are reported in Table 2 with their 95% confidence intervals (CIs). The estimated component-wise IDEs are shown in Figure 5.

Table 2: The estimated DE and TIDE of fat intakes on BMI through the gut microbiome using the proposed compositional mediation model (CMM). The values in parenthesis represent 95% CIs. The second row presents the percentages of DE and TIDE in the total effect of fat intakes on BMI.

DE	TIDE
0.949 (0.003, 1.901)	0.732 (-0.331, 2.114)
56.5%	43.5%

The estimated DE and TIDE are 0.949 and 0.732, respectively, which correspond to 56.5% and 43.5% of the total effect of fat intakes on BMI. Note that under LSEM the total effect of treatment can be decomposed into direct and indirect effects. The estimated DE is statistically significant at the significance level of 0.05, but the estimated TIDE, as well as all component-wise IDEs, are not. It is most likely due to an insufficient sample size. In the simulation, we observed that TIDE of 0.75 has a significant probability of its CI containing zero, and component-wise IDEs at around 0.2 can be rarely detected with a sample size of 100.

As shown in Figure 5, the potential genera that would have statistically significant non-zero component-wise IDEs with a sufficient sample size are *Alistipes, Oscillibacter, Acidaminococcus*, and *Allisonella*. All these genera have positive values for IDEs, but their responses to fat intakes are quite different. The abundance of *Alistipes* and *Oscillibacter* is negatively correlated with fat intake and BMI, whereas that of *Acidaminococcus* and *Allisonella* is positively correlated with fat intake and BMI. Based on this information, we can hypothesize that the increase in fat intake causes the decrease in the abundance of *Alistipes* and *Oscillibacter* and the increase in the abundance of *Acidaminococcus* and *Allisonella* which in turn cause the increase in BMI. Lam *et al.* (2012) identified Oscillibacter-like organisms as a potentially important gut microbe that mediates high fat-induced gut dysfunction and gut permeability and showed that decrease of *Oscillibacter* led to increased gut permeability, which was shown to be associated with obesity (Teixeira *et al.*, 2012). These observations were largely consistent with what were observed in mice fed with high-fat diet (Daniel *et al.*, 2014).

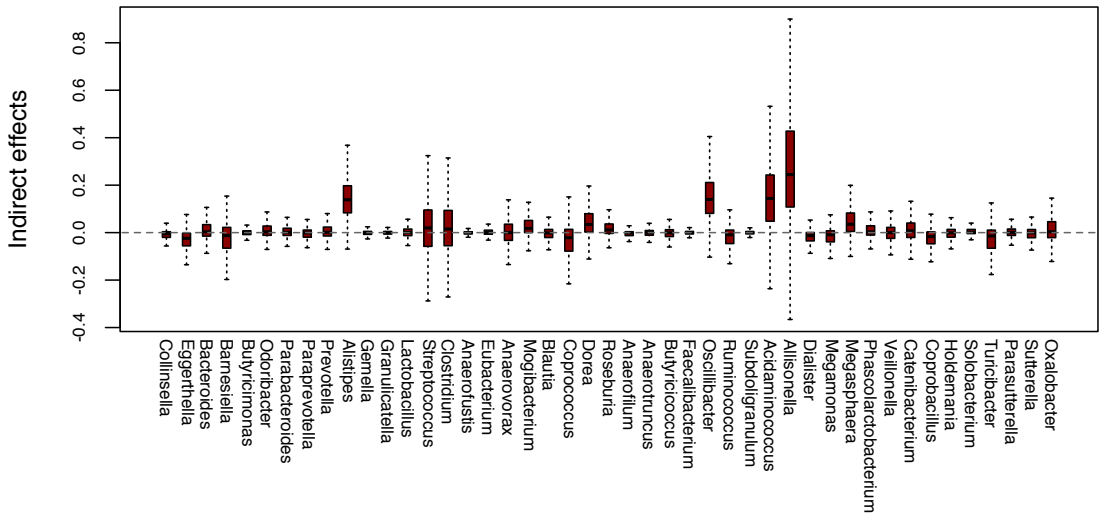

Figure 5: The estimated component-wise IDEs of fat intakes on BMI through the gut microbiome.

6 Discussion

In this study, we propose a compositional mediation model for a continuous outcome. Our method takes the characteristics of compositional data into account and treats the whole compositional mediators as a unit, that is, it estimates the effect of treatment on compositional mediators simultaneously instead of each mediator separately. In our simulation studies, we have shown better performance of our method over the two potential methods for compositional mediators. Our method also provides a clear interpretation of component-wise indirect effects. Its application to gut microbiome and BMI dataset has clearly indicated a potential mediation effect of the gut microbiome in linking the association between fat intake and BMI. Although the results were not statistically significant, several bacterial genera have been shown to be directly associated with gut permeability and therefore BMI. It would be interesting and important to replicate these results in larger datasets.

In many clinical microbiome studies, the outcome variable is binary such as whether a subject is diseased or not. In these cases, Model (2) can be rewritten for logistic or probit regression, assuming the outcome variable is a latent continuous variable indicated by an observed dichotomous variable. Then, Model (1) and the modified Model (2) will provide the identifiability of direct and indirect effects (Winship and Mare, 1983). Another interesting extension of our method is for longitudinal compositional data, which is also very common in microbial studies.

Even though we used only a continuous treatment variable in the simulation study, a binary treatment

variable can be used without any modification in our method. However, a treatment variable with more than two categories cannot be used since it requires $k-1$ parameters to fully represent the effect of a categorical variable with k mutually exclusive categories on a dependent variable. A compositional mediation model with a general categorical treatment variable is another interesting extension of our method.

Appendix A Proof of Theorem 1

Given $Y \perp\!\!\!\perp M_j|(\mathcal{T}, X)$, $Y \perp\!\!\!\perp M_j^c|(\mathcal{T}, X)$ since

$$\begin{aligned} p(y \cap m_j^c|\tau, x) &= p(y|\tau, x) - p(y \cap m_j|\tau, x) = p(y|\tau, x) - p(y|\tau, x)p(m_j|\tau, x) \\ &= p(y|\tau, x)\left[1 - p(m_j|\tau, x)\right] = p(y|\tau, x)p(m_j^c|\tau, x). \end{aligned}$$

Since $Y \perp\!\!\!\perp M_j^c|(\mathcal{T}, X)$ and $M_i, \forall i \neq j$ are disjoint,

$$\begin{aligned} p\left(y \cap \bigcup_{i\neq j}^{k} m_i \middle| \tau, x\right) &= p(y|\tau, x)p\left(\bigcup_{i\neq j}^{k} m_i \middle| \tau, x\right) = p(y|\tau, x)\left[\sum_{i\neq j}^{k} p(m_i|\tau, x)\right] \\ &= \sum_{i\neq j}^{k} p(y|\tau, x)p(m_i|\tau, x). \end{aligned} \tag{14}$$

By the distributive law, we can also have

$$\begin{aligned} p\left(y \cap \bigcup_{i\neq j}^{k} m_i \middle| \tau, x\right) &= p\left(\bigcup_{i\neq j}^{k} (y \cap m_i) \middle| \tau, x\right) \\ &= \sum_{i\neq j}^{k} p\left(y \cap m_i|\tau, x\right). \end{aligned} \tag{15}$$

To satisfy the equality of Equations (14) and (15) for any y and $m_i, \forall i \neq j$ satisfying $Y \perp\!\!\!\perp M_j^c|(\mathcal{T}, X)$, $p(y \cap m_i|\tau, x) = p(y|\tau, x)p(m_i|\tau, x), \forall i \neq j$, that is, $Y \perp\!\!\!\perp M_i|(\mathcal{T}, X), \forall i \neq j$. □

Appendix B Proof of Theorem 2

It has been shown that under the assumption (3) - (4), direct and indirect effects are identifiable and equal to c and $(\log A)^T B$, respectively, within the framework of LSEM (Imai *et al.*, 2010). Thus, showing the proposed compositional mediation model falls into LSEM suffices to prove Theorem 1. Applying the additive logratio

transform to both sides of Model (1), we have

$$\log \frac{M_{ij}}{M_{ik}} = \log \frac{M_{0j}}{M_{0k}} + \tau_i^* \log \frac{A_j}{A_k} + \epsilon_{ij}, \quad \forall j \in \{1, 2, \ldots, k-1\}, \tag{16}$$

where $\epsilon_{ij} = \log(E_{ij}/E_{ik})$. Because of the unit sum constraint of Model (2), we can write Model (2) as

$$\begin{aligned} y_i &= c_0 + \tau_i^* c + (\log M_{i1}) B_1 + \cdots + (\log M_{i(k-1)}) B_{k-1} \\ &\quad - (\log M_{ik})(B_1 + \cdots + B_{k-1}) + \varepsilon_i \\ &= c_0 + \tau_i^* c + \left(\log \frac{M_{i1}}{M_{ik}}\right) B_1 + \cdots + \left(\log \frac{M_{i(k-1)}}{M_{ik}}\right) B_{k-1} + \varepsilon_i. \end{aligned} \tag{17}$$

Let $g_{ij} \equiv \log \frac{M_{ij}}{M_{ik}}$ and $f_j \equiv \log \frac{A_j}{A_k}$, then we have

$$\begin{aligned} g_{i1} &= g_{01} + \tau_i^* f_1 + \epsilon_{i1}, \\ g_{i2} &= g_{02} + \tau_i^* f_2 + \epsilon_{i2}, \\ &\vdots \\ g_{i,k-1} &= g_{0,k-1} + \tau_i^* f_{k-1} + \epsilon_{i,k-1}, \end{aligned}$$

and

$$y_i = c_0 + \tau_i^* c + g_{i1} B_1 + \cdots + g_{i,k-1} B_{k-1} + \varepsilon_i.$$

Therefore, Equations (16) - (17) belongs to LSEM. □

Appendix C Model with confounding variables

Let X be a covariate and Ψ be a vector of its parameter. Then, the parameters A and Ψ can be estimated by solving

$$\begin{aligned} (\hat{A}, \hat{\Psi}) &= \underset{A, \Psi, M_0 \in \nabla^{k-1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| M_i \ominus M_0 \oplus A^{\tau_i^*} \oplus \Psi^{x_i^*} \right\|^2 \\ &= \underset{A, \Psi, M_0 \in \nabla^{k-1}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{k-1} \left\{ (k-1) \left[\log \left(\frac{M_{ij} M_{0k} A_k^{\tau_i^*} \Psi_k^{x_i^*}}{M_{ik} M_{0j} A_j^{\tau_i^*} \Psi_j^{x_i^*}} \right) \right]^2 \right. \\ &\qquad \left. - \log \left(\frac{M_{ij} M_{0k} A_k^{\tau_i^*} \Psi_k^{x_i^*}}{M_{ik} M_{0j} A_j^{\tau_i^*} \Psi_j^{x_i^*}} \right) \sum_{\ell \neq j}^{k-1} \log \left(\frac{M_{i\ell} M_{0k} A_k^{\tau_i^*} \Psi_k^{x_i^*}}{M_{ik} M_{0\ell} A_\ell^{\tau_i^*} \Psi_\ell^{x_i^*}} \right) \right\}, \end{aligned}$$

or solving the system of $2(k-1)$ linear equations:

$$\begin{bmatrix} D(\tau^{*2}) & D(x^*\tau^*) \\ D(x^*\tau^*) & D(x^{*2}) \end{bmatrix} \begin{bmatrix} f_\alpha \\ f_\phi \end{bmatrix} = \begin{bmatrix} m(\tau^*) \\ m(x^*) \end{bmatrix}$$

with a constraint $A^T\mathbf{1}_k = 1$ and $\Phi^T\mathbf{1}_k = 1$, where

$$D(\xi) = \begin{bmatrix} (k-1)\sum_{i=1}^n \xi_i & -\sum_{i=1}^n \xi_i & \dots & -\sum_{i=1}^n \xi_i \\ -\sum_{i=1}^n \xi_i & (k-1)\sum_{i=1}^n \xi_i & \dots & -\sum_{i=1}^n \xi_i \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{i=1}^n \xi_i & -\sum_{i=1}^n \xi_i & \dots & (k-1)\sum_{i=1}^n \xi_i \end{bmatrix},$$

$$f_\alpha = \left[\log(A_1/A_k), \log(A_2/A_k), \dots, \log(A_{k-1}/A_k)\right]^T,$$

$$f_\phi = \left[\log(\Phi_1/\Phi_k), \log(\Phi_2/\Phi_k), \dots, \log(\Phi_{k-1}/\Phi_k)\right]^T,$$

$$m(\xi)_j = k\sum_{i=1}^n \xi_i \log M_{ij} - \sum_{i=1}^n \sum_{\ell=1}^k \xi_i \log M_{i\ell},\ \forall j \in \{1, \dots k-1\}.$$

Similarly, the estimate of the parameter A with additional covariates can be obtained as long as the covariates are not highly correlated. The number of linear equations to be solved with n_c covariates is $n_c(k-1)$.

Appendix D Variance of indirect effects

To approximate the variances of a total indirect effect and component-wise indirect effects, the first order asymptotic method described in (Bollen, 1987) is used. That is, the variance of each component-wise indirect effect is approximated by

$$\mathrm{Var}(\log(k\hat{A}_j)\hat{B}_j) \approx [\mathbb{E}(\log(k\hat{A}_j))]^2\mathrm{Var}(\hat{B}_j) + [\mathbb{E}(\hat{B}_j)]^2\mathrm{Var}(\log(k\hat{A}_j))$$

and the variance of a total indirect effect is approximated by

$$\begin{aligned}\mathrm{Var}((\log \hat{A})^T \hat{B}) \approx & \sum_{j=1}^{k} [\mathbb{E}(\log \hat{A}_j)]^2 \mathrm{Var}(\hat{B}_j) + \sum_{j=1}^{k} [\mathbb{E}(\hat{B}_j)]^2 \mathrm{Var}(\log \hat{A}_j) \\ & + 2 \sum_{j<\ell}^{k} \mathrm{Cov}(\log \hat{A}_j, \log \hat{A}_\ell) \mathbb{E}(\hat{B}_j) \mathbb{E}(\hat{B}_\ell) \\ & + 2 \sum_{j<\ell}^{k} \mathrm{Cov}(\hat{B}_j, \hat{B}_\ell) \mathbb{E}(\log \hat{A}_j) \mathbb{E}(\log \hat{A}_\ell).\end{aligned}$$

References

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc., Ser. B (Stat. Methodol.)*, **44**(2),139-177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. New York: Chapman & Hall.

Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, **71**(2), 323-330.

Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, **29**(4), 1165-1188.

Billheimer, D., Guttorp, P. and Fagan, W. F. (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association*, **96**(456), 1205-1214.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*. **17**, 37-69.

Bray, G. A. and Popkin, B. M. (1998). Dietary fat intake does affect obesity! *Am J Clin Nutr*. **68**(6), 1157-1173.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Method, Theory and Applications*. Berlin: Springer.

Chén, O. Y., Crainiceanu, C. M., Ogburn, E. L., Caffo, B. S., Wager, T. D. and Lindquist, M. A. (2015). High-dimensional Multivariate Mediation with Application to Neuroimaging Data. arXiv:1511.09354

Daniel, H., Gholami, A. M., Berry, D., Desmarchelier, C., Hahne, H., Loh, G., Mondot, S., Lepage, P., Rothballer, M., Walker, A., Böhm C., Wenning, M., Wagner, M., Blaut, M., Schmitt-Kopplin, P., Kuster,

B., Haller, D. and Clavel, T. (2014). High-fat diet alters gut microbiota physiology in mice. *ISEM J*, **8**(2), 295-308.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Huang, Y. T. and Pan, W. C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, **72**(2), 402-413

Lam, Y. Y., Ha, C. W., Campbell, C. R., Mitchell, A. J., Dinudom, A., Oscarsson, J., Cook, D. I., Hunt, N. H., Caterson, I. D., Holmes, A. J. and Storlien, L. H. (2012). Increased Gut Permeability and Microbiota Change Associate with Mesenteric Fat Inflammation and Metabolic Dysfunction in Diet-Induced Obese Mice. *PLoS ONE*, **7**(3): e34233.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, **25**(1), 51-71.

Imai, K., Keele, L. and Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychol Methods*, **15**(4), 309-334.

Machado, J. A. F. and Parente, P. (2005). Bootstrap estimation of covariance matrices via the percentile method. *The Econometrics Journal*, **8**(1), 70-78.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, **7**(1), 83-104.

Ley, R. E., Turnbaugh, P. J., Klein, S. and Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, **444**, 1022-1023.

Lin, W., Shi, P., Feng, R. and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, **101**(4), 785-797.

Pearl, J. (2001). Direct and indirect effects. *In Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 411-420. San Francisco, CA: Morgan Kaufmann.

Preacher, K. J. and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav res methods*, **40**(3), 879-891.

Rubin D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*. **100**(469), 322-331.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *Annals of Mathematical Statistics*. **21**(1), 124-127.

Shi, P., Zhang, A. and Li, H. (2016). Regression Analysis for Microbiome Compositional Data. *Annals of Applied Statistics*, **10(2)**, 1019-1040.

Shrout, P. E. and Bolger, N. (2002). Mediation in Experimental and Nonexperimental Studies: New Procedures and Recommendations. *Psychological Methods*, **7**(4), 422-445.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, **13**, 290-312.

Teixeira, T. F., Collado, M. C., Ferreira, C. L., Bressan, J. and Peluzio, M. C. (2012). Potential mechanisms for the emerging link between obesity and increased intestinal permeability. *Nutr Res.*, **32**(9), 637-47.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R. and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027-1031.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology* **172**(12), 1339-1348.

VanderWeele, T. J. and Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiol Method*. **2**(1): 95-115.

Winship, C. and Mare, R. D. (1983). Structural Equations and Path Analysis for Discrete Data. *The American Journal of Sociology*. **89**(1):54-110.

Wu, G., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D. and Lewis J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*. **334**(6052) 105-108.

Zhao, Y. and Luo, X. (2016). Pathway Lasso: Estimate and Select Sparse Mediation Pathways with High Dimensional Mediators. arXiv:1603.07749.